INSTITUT POUR LES ETUDES
SEMANTIQUES ET COGNITIVES

UNIVERSITE DE GENEVE

A PRELIMINARY STUDY ON LINGUISTIC IMPLICATIONS OF RESOURCE CONTROL IN NATURAL LANGUAGE UNDERSTANDING

J. S. Bien

1980

No. 44

WORKING PAPERS

Fondazione Dalle Molle

Acknowledgement

The ideas expounded in this paper were worked out in the author's own institution. A brief stay at ISSCO allowed their formulation on paper.

The author would also like to thank Yorick Wilks and Boguslaw Jankowski for their interest in this work and for much helpful discussion.

J. S. Bien
Institute of Informatics
University of Warsaw
P.O. Box 1210,
00-901 Warsaw
Poland

1. INTRODUCTION

The aim of this paper is to formulate some hypotheses concerning the organisation of language processing by human and computer. The ideas advocated here are not yet developed enough to warrant a detailed presentation. The author has decided to present them to the scientific community despite the early stage of the inquiry because of their possible impact, especially for the investigations of the up-to-now completely mysterious fact that English articles are rendered in Slavonic languages mainly by word order and vice versa; explaining the phenomenon, viewed by the author as different ways of controlling the depth of nominal phrase processing, would serve both to improve our understanding of the nature of language and assist the design of reliable machine translation systems. The paper is intended to be selfcontained, but familiarity with the work of Norman and Bobrow (1975,1976,1979, Bobrow and Norman 1975) is strongly recommended. I would like also to comment briefly on the relation of the present paper to the multiple environments model of natural language, formulated first in (Bien 1975).

The basic assumption of the multiple environments approach is that if a computer system is to understand natural language fully, it must be a sufficiently adequate model of the relevant cognitive processes of human beings. Therefore it is not enough to work out separately the problems of language understanding (e.g. pronoun resolution); the solutions should be integrated into a proper overall structure. In the first phase of my research (Bien 1976, 1976a) my attention was concentrated on modelling by means of "environments" the division of memory into different ontological classes (e.g. addressee's beliefs, the addressee's beliefs about the sender's beliefs about the addressee, etc.). A similar approach was advocated independently by Cohen (1978) and used successfully by his colleagues (Perrault, Allen, Cohen, 1978). Although the notion of environments and its equivalents proved useful, its psychological reality remains still an open question (Bien 1977, Wilks and Bien 1979).

The multiple environments model was presented using «utterances as programs» metaphor (Davies and Isard 1972, Longuet-Higgins 1972), which made it rather hard to understand for readers without a computer science background. On the other hand, the metaphor served different (although not contradictory) purposes than e.g. in (Miller, Johnson-Laird 1976). Instead of looking for the analogues of compilation and execution, I insisted that the language processor does not make any decision of his own, just as a computer hardware or software interpreter executes step by step its current program. In consequence, I claimed that the processor can process properly only an utterance formulated according to some specific "minimal effort principles" (Bien 1977). The principles allow the sender to encode (usually subconsciously) in the structure of the utterance some control information for the language processor of the addressee. In the present paper we will see what such control information might look like.

For some readers the following example may serve as an illustration of the approach advocated here. The human cognitive processor has some intrinsic limitations, just as the size of relevant registers limits the precision of floating point arithmetic. A typical compiler generates the binary code in a straightforward fashion, usually preserving the source code order of operations; therefore we can specify the operations on the source code level in such an order as to prefer e.g. adding numbers of similar sizes (which results in minimization of the round-off errors introduced by the normalisation of the components). In other words, we do have a way of controlling the precision of floating point arithmetic although at first glance it seems uncontrollable. Just as the order of components to be added should fit the implicit rules of the computer system used, the utterance, to be properly understood, must be adapted to the relevant features of the human language processor.

The limits of the paper do not permit a full discussion of the pros and cons for the opinions presented; in particular, it is not possible to give proper credit to all the works which, influencing my way of thinking about language, have led me more or less directly to the formulation of the "resource control hypothesis".

2. BASIC FEATURES OF THE COGNITIVE PROCESSOR

We assume that the processor's behaviour is determined first of all by two properties formulated by Bobrow and Norman (1975):

"There is a limit to the processing resources available to the organism" (p. 140)

which states the resource limit principle. It is very useful to explain the interaction between various mental processes, classified as resource-limited (when an increase in the amount of processing resources results in improved performance) or data-limited (when performance is independent of processing resources). The second property, called here the input assimilation principle and originally formulated as "all data are to be accounted for" (p. 140), states that for all sensory data "some conceptual scheme must be found for which these data are appropriate" (p. 144). In other words, the cognitive processor permanently monitors the perceived information and attempts to fit it into its expectations.

We assume also two other principles : "steady availability of resources" and "steady consumption of resources". The processing power of the processor changes relatively slowly and in rather narrow bounds, being probably a function of the organism arousal; on the other hand, the processor is never idle. Whenever the processing of the external data (which has the highest priority because of the input assimilation principle) doesn't require all the resources, the available power is used for some spontaneous action of the processor. We assume that such a spontaneous action consists almost exclusively in resource-limited processes like memory search or inferences.

We take for granted that processing is performed in parallel. In consequence, we imagine the cognitive processor as a network of numerous microprocessors; the pseudo-physical properties of the network constitute an intrinsic part of the model of the cognitive system to be advocated here. According to our belief that the processor doesn't take any sophisticated decisions of its own, we assume that the microprocessors are supervised according to some very simple principle. Its first approximation may be e.g. the following rule : if a computation branches into n alternatives, demand n microprocessors; if they are not available, suspend the computation. A process, even a suspended one, may receive partial output from other related processes. Therefore the computation is resumed either because the required number of microprocessors become available, or because the output received caused the computation to reduce its demand.

On the other hand, it may happen that two processes request the same particular microprocessor; it seems natural to prefer the process which is more straightforward, i.e. involves less alternative computations and, in consequence, demands fewer microprocessors. In what follows we will call such processes more deterministic, so this

allocation rule can be stated as the preference for deterministic computation rule. The rule, in turn, can be viewed as a new kind of minimal effort principle, because the statement "deterministic processes are easier to compute" is intuitively true.

We expect that the total number of demands always exceeds the resources available; the problem of deadlock doesn't exist here, because it is intuitively acceptable that some demands of the lowest priority may never be satisfied.

Because of the large number of microprocessors, it would be unnatural to expect that every microprocessor communicates equally easily with all others. On the contrary, it seems that a microprocessor should communicate directly only with its neighbours in the network, and the messages to other microprocessors must be transmitted indirectly. In consequence, the flow of computation may often depend on the timing of input/output messages, determined by the pseudo-physical configuration of micro-processors running the interacting processes under consideration.

The strict definition of the processor advocated here is a very difficult task. A promising direction of research may be «augmented logic programming», i.e. logic programming (Kowalski 1979) which provides simplicity, parallelism, continuous output of partial results and other interesting features, supplemented by some structuralisation in the style of actors (Hewitt, Bishop and Steiger 1973) with the control strategy proposed above; some valuable suggestions may be perhaps also supplied by neurophysiology. It would be interesting, for example, to investigate the cerebral blood flow to verify whether the oxygen supply might be the primary factor limiting the processing power.

It should be noted that in the above discussion we have in mind exclusively subconscious processes. Contrary to Bobrow and Norman, I have no strong intuitions about the purpose of consciousness. For example, the "processing symmetry principle" stating that "the processing can consist of fitting input to expectation" or "finding structures in which to embed the input" (Bobrow and Norman 1975 : 140) can be viewed not as an intervention of the central mechanism (1975 : 146), but just as a side-effect of the resource allocation strategy.

3. SOME PROPERTIES OF COGNITIVE MEMORY

The processor outlined in the preceding paragraph has an interesting feature : it processes data according to simple rules but it is very sensitive to the context of computation, consisting in the physical configuration of processes. As before, we assume that the way the configuration changes over time must be governed by some very primitive rules. The centrifuge metaphor, described below, is a first attempt to approximate the rules searched for.

Let us imagine a centrifuge, preferably with a sphere-like container, and consider the fluid particles to be the analogues of memory schemata in the spirit of Bobrow and Norman (1975). The surface of the fluid, characteristically shaped, models the focus of attention. The centrifugal and gravity forces cause the particles to move away from the surface, some of them even hit the walls of the container; by analogy, we attribute to the memory schemata the tendency to be forgotten, realised by physically transmitting the information which constitutes the schemata under consideration to the microprocessors situated more and more distantly from the focus of attention; these schemata which "hit the container" are just overriden by some incoming sensory data (the memory is finite, so every new information has to erase physically some old one).

Before we explain why everything is not always forgotten, it is necessary to describe the structure of memory schemata. They consist in components called slots in artificial intelligence jargon, containing context-dependent descriptions (Norman, Bobrow 1979). The descriptions serve two purposes : they yield partial information about the components, and specify the criteria for recognising, in the particular context, a possible component instance in the memory. The components of a schema are usually interconnected, and matching a component description against the schema it refers to results in propagating to other components the information constituted by the instantiation, established during the matching operation, of the variables in the context dependent description under consideration. The values received by the variables consist generally in other context-dependent descriptions which may in turn contain new variables; however, every instantiation of a variable increases the amount of information stored in the schema. In consequence, we may compare the number of variable instantiations in differen schemata and so call the schemata more or less instantiated.

If a schema is successfully matched against a context-dependent description, the transmission of information is bidirectional in the general case : from the instantiating schema to the instantiated one (i.e. the schema to which the instantiated description belongs to) and vice versa (cf. for example the unification operation in logic programming). In a special case it may happen that, due mainly to the complexity of the relations between a schema's components, the instantiation of a description and the propagation of the variable setting to the other components consists in sophisticated processing of input information.

We assume that whenever a schema is in the focus of attention area (the surface of memory), an attempt is made to instantiate all its components. The proper queries are sent step by step to more and more microprocessors; the query transmission, which takes a considerable amount of time because of the postulated pseudo-physical properties of the network, has the character of a wave going through the whole memory (possibly interfering with other such waves initiated by other components of the schema or other schemata present in the focus area). To stress these facts we will refer to a query transmission as broadcasting.

Whether the schema reaches the surface or not depends on what priority relative to other responding schema is attributed to it by the scheduling rules; in any case, the copy of the schema is closer to the surface than the original schema was; although it is now subject to the general rules of forgetting, the memory of information stored in it was refreshed. The centrifuge analogue of the situation is when a particle does not hit the container, but is brought by the undercurrent to the central whirl or to the surface (if we identify the focus of attention with the consciousness, we call the whirl "subconsciousness").

The focus of attention is a region of the memory where an important amount of processing capacity is concentrated. It is used mainly for processing schemata on the component level, e.g. transmitting information between components belonging to one or to several schemata. in other parts of the memory the processing capacity available allows only processing of the most prominent components of a schema; such components may be called headers or clues. Only the components processed decide whether a schema will respond to a query. It should be stressed that the focus of attention region includes also purely sensory schemata (the fluid surface in the centrifuge borders with the walls of the container, which are intepreted as the source of the sensory data to be monitored.

As may be expected, we do not postulate different kinds of memory. The differences in memory operation are attributed to the structure of individual schemata (semantic memory schemata have different kinds of components and clues to episodic schemata) or to the distance of the schemata from the focus of attention (short versus long term memory, etc.).

The cognitive system advocated here (the processor and the memory) shows a very important feature : although based on simple deterministic rules, it can react differently to identical stimuli because of the ability to memorise, again by simple deterministic rules, the history of previous computations in the form of physical configuration of the memory schemata. Its enormous possibilities are most easily demonstrated by recalling the "intelligent" behaviour of the homeostat (Ashby 1952), which was an automaton of only 390, 625 states; if the cognitive system is treated as a black box automaton,

then its number of states is incomparably greater, so we can expect from it respectively more sophisticated behaviour. On the other hand, the cognitive system has many more parameters which can be changed independently than does the homeostat, so the chances of tuning the system to the desired kind of behaviour are much better.

Of course, the ideas presented do not explain all the known phenomena, but for the time being it is sufficient that, as far as I know, the centrifuge hypothesis does not contradict any. It may also be the case that the fluid particles in a centrifuge circulate in a different way than described (I am not an expert in fluid mechanics), but this would be obviously irrelevant to our discussion.

4. MEMORY SCHEMATA

In this section we will develop further the centrifuge model of memory by discussing the life cycle of a typical schema.

Depending on the current direction of processing, a physical instance of a prototype schema is either already present in the focus area, or the prototype is searched for. In any case the schema we will discuss is created in the focus area by merging the prototype with the input data which are to be accounted for; the input may consist in the real input of sensory data or in some output of other cognitive processes. Depending on the centrality of the schema position in the focus area, more or less intensive attempts are made to instantiate the context-dependent descriptions contained in the components of the schema.

It is natural to assume that one of the schema's components points to the schema's prototype. Because the attempt is made to instantiate all the components, it ensures that the prototypes are not forgotten and that the ones most frequently used are kept close to the focus.

While the queries for the slot instances are broadcasted, the schema itself is already subject to the tendency of being removed from its current position to a less central one. A schema which is really useful for the processing in progress is highly interconnected with other schemata in the focus, the interconnections actually consisting in transmitting the information, so that useful schemata should be relatively more instantiated than the less useful ones. A more fully instantiated schema demands a more deterministic processing; because the deterministic computations are preferred by the resource allocation rule, such schema will get the resources requested at the cost of the less fully instantiated ones. When free storage is needed in the focus for a new schema, the useful schema will use the resources to stay in the focus area, but the other schemata will be removed to the region where they can be maintained with smaller resource consumption, i.e. outside the focus area. To make place for them, the schemata residing close to the focus area will be transmitted to less central positions, etc.

Let us assume that the schema under consideration is a useful one and has resisted the attempts to remove it from the focus area. In the meantime some schemata matching the context-dependent descriptions of its components were found and transmitted toward the focus area. Those which looked good enough to the resource allocation routine got their places in the focus area, the others failed and reside more or less closely to the focus. Additionally, some schemata may be still in the course of transition toward the focus, and some relevant schemata may not be reached yet by the broadcast of the inquiries. The schemata in the focus area instantiated the respective components of the schema under discussion, i.e. the information was exchanged between the instantiated and instantiating schemata. It may happen

now that some other schema is more fully instantiated than the one under discussion, either due to the direction of the information propagation, or because new input data fit the other schema well enough. Let us assume then that our schema cannot resist this time the new attempts to remove it, and becomes forgotten.

In the general case the situation is slightly more complicated. When the schemata are on the surface, the borders between them practically do not exist, because the concentration of the processing capacity is so high that every component gets the resources independently of the schema to which it belongs. When a new schema arrives at the surface, room is made for it by looking for the components which are less useful. When they are found, they are packed together in such a way that as few links to other schema are cut as possible : usually the interconnections between the components of a schema are denser than the interconnection between the components belonging to different schemata, so the procedure results in reconstructing the schema brought to the surface. Nevertheless, in the general case a new schema can be created and subjected to the forgetting process.

The fate of the schema removed from the focus area depends now on two factors : gravity and centrifugal forces. The gravity force causes the schemata to drown, i.e. to be transmitted to the less active regions of the memory, where their maintenance consumes less resources. Economy of resources is gained by processing only the most prominent components, therefore it may happen that a schema will not respond to a relevant query merely because it is referred to by the description of a less prominent component. Actually the gravity force is a result of the fact that all the schemata are pushed down and down to make place for the schemata removed from the focus area. A natural assumption is that with growing depths a decreasing number of components is subject to processing.

On the very bottom of the memory, the processing capacity is so low that the query cannot be compared even with one component of every schema, but a component of a schema is to be chosen more or less randomly for comparison. From the point of view of the query sender, the schemata totally ignored during the comparison are represented by the component which happened to be processed. In the general case the component may be unrelated to the schemata skipped, and such misrepresented schemata can be brought to the focus area only accidentally. It might be the case that e.g. the memories of early childhood which last for years and come to mind quite spontaneously can be explained just this way.

Let us imagine now that a query was broadcast to which our schema is relevant. When it arrives at the schema, it is matched against those components of it, which are subject to processing at the given depth. If the answer is positive, the schema moves through the undercurrent towards the surface, i.e. it is transmitted by recopying, through channels different from those used for broadcast, to a more and more

central position. When approaching the surface, the channels become more and more busy and the schema has to compete with others on a first come - first served basis. The schemata waiting for channels to be free are the analogue of the whirl, and their requests for storage constitute the centrifugal force mentioned above. If a schema is pushed by its competitors far enough from the whirl, it is not able to recover its status and become subject again to the gravity and centrifugal forces.

Transmitting by recopying is subject to various distortions : some information may be garbled by hardware errors or the transmission cannot be finished because of lack of resources.

The first case is rather uninteresting, although it can serve as an explanation of some deformations of our memories. The second case is a very important one. If we assume that the schema is transmitted approximately in the top-down order of its internal pointer structure, then the interrupted transmission may result in a syntactically correct schema which lacks only some most specific information from the original one. Under proper circumstances such a transmission may create a new prototype schema. This fact, together with the possibility mentioned above of creating new schemata in the focus area, might constitute the essence of some learning abilities.

When a schema is copied, two eventualities arise. If there is a pending request for storage, the original instance of the schema is erased by its more central neighbour. If not, the very instance of the schema is left on its place and is subject to general rules of forgetting : it will be drowning and pushed by the centrifugal force. At the end the schema will reach the walls of the container, i.e. it will be erased and the storage space formerly occupied by it will be used for sensory data to be processed. It should be noted that this does not mean that the information is completely forgotten, because some other copies of the schema may still exist.

As can be seen from above, retrieving old information from memory is very resource consuming. Introducing a new object into memory consists just in creating a proper context-depending description, which is a data-limited process requiring a very small amount of resources. On the other hand, making the new object available for future retrievals consists in relating it to proper cues, which is again a resource-limited process with heavy demands on processing power.

5. SOME LANGUAGE UNIVERSALS

We will limit our attention to the problem of understanding natural language, because it seems to be more basic than language generation. Passive knowledge of language is always better than active; this is also true for all the stages of (both first and second) language acquisition, it is natural to assume that the language universals are determined mainly by the properties of the understanding process. Additionally, it is quite probable that in careful speech the generation process is guided by re-analysis of the output.

We take for granted that different aspects of an utterance to be understood are processed in parallel. Following Bien (1976) we will distinguish four main parallel processes (called levels in the sequel) : sorbtion, syntactics, semantics and assimilation. Sorbtion consists in the acoustic or optical recognition of utterance elements. It is a data-limited process and has the highest priority because of the data assimilation principle. The function of the second level is relatively close to the traditional notion of syntax: to recognise the clues consisting e.g. in inflexional endings and some other surface phenomena for determining how to assemble the meanings of the utterance elements (to tell which meanings go together in the semantic structure of the utterance, it is not necessary to know what these meanings actually are). For the time being, we imagine the process of syntactics to resemble that of Marcus' deterministic parser (Shipman, Marcus 1979). The syntactics is also a data-limited process, but it should be noted that its demands for resources may vary relatively much, depending on the current stage of the parsing process.

The third level, the semantics, is responsible for two tasks. First, its job is to construct the complex of memory schemata representing the meaning of the utterance processed; some of its operations can perhaps be done in parallel in the style of Smith and Rawson (1976). Secondly, it should instantiate the proper parts of the semantic representation schemata with the old information stored in the memory. Because of the second task the semantic level as a whole is a resource-limited process; however, the amount of processing actually performed by it depends strongly on the resource allocation preference for a deterministic flow of computations.

The fourth level, the assimilation, consists in spontaneous processing not related directly to the literal meaning of the utterance, e.g. inferring the connections between the utterance meaning and other known facts, which supplies it with useful clues for future retrievals, but does not belong to the meaning itself. The assimilation is also a resource-limited process, and its priority is even lower than that of the semantics, because it is very loosely related to the sensory data.

The input to the syntactic level is of course the output of sorbtion, the input to the semantic level is the output of the syntactics, etc. We assume however that all the processing is done by some memory schemata, and e.g. the semantics can influence the processing on other levels by changing the saliency of the relevant schemata.

Both in phylogenesis (the emergence of language in humans) and in ontogenesis (language acquisition by individuals), it is the spoken language which appears first. In consequence, the language understanding mechanisms are developed under the constraint of real-time processing; we consider the fact to be of crucial importance for the explanation of the organisational principles of language processing. The cognitive processor is so adapted to real-time language understanding that it probably processes written texts in the same manner, the only difference consisting in continuous supply of data by the optical perception processes instead of the acoustic ones (everybody has experienced the situation when his eyes scan a text while his attention has been drawn away; although the eye movement can be avoided or minimalised by a special training in fast reading techniques, which increases the throughput of the optical processing, the output of recognition is probably still supplied sequentially).

According to the naive view, language communication is an easy task for humans. On the other hand, the participants in language communication are able to perform simultaneously only highly automated activities like walking or driving etc. (unless the communication itself is not a stereotyped one). This fact can be explained by assuming that language understanding requires almost full capacity of the cognitive processor, so it interferes with other activities except those which consume a negligible amount of resources; in what follows this assumption is called the full capacity principle.

The automatic activities are assumed to be just very economical as far as resource consumption is concerned, because we expect them to be characterised by very high saliency of the relevant schemata, an optimal organisation of information into the schemata, and an efficient way of referencing the related schemata (by not using the context-dependency unless necessary).

All the factors minimize the need for resource-consuming search of memory; the first one is decisive for temporary skills which are easily forgotten, the others enter into consideration only for heavily practised and well memorised skills.

As a consequence of the assumption that processing of utterances is done almost on the verge of the processor capacities, we expect that different subprocesses of the understanding process are all the time competing for resources. This solves in a natural way the old problem of when to stop spontaneous inferences; because of their low priority relative to other levels of processing, they stop simply when they exhaust the resources available.

Another consequence of the full capacity principle is the expectation that often the processing resources happen to be insufficient to understand an utterance in real-time. We think it is really the case, as it was noted already by Ziff : "Sometimes some of us understand some of what is said" (1972). Our claim is that if during the processing of an utterance, some of demands for resources cannot be fulfilled, the utterance is found by the hearer to be more or less awkward. Actually what happens in such a situation is that the processing is more or less serialised and understanding of the sentence takes longer time; if the additional amount of time needed is small relative e.g. to the intersentence gap, the sentence is understood and the processor is able to catch up the input data at the cost of e.g. some spontaneous inferences. If the amount of time necessary to recover from the processor overload is greater, it may lead to unsolvable conflicts about what should be processed in a given moment (just one of the competing tasks must then be aborted).

A special case of processing serialisation is where a sentence is fully ambiguous, i.e. in the given context two interpretations are on equal rights and processing of both alternatives in parallel would cause an important overload of the processor. In such a situation the serialised subprocesses of understanding must be run at relatively long time intervals and cannot use common internal data; in consequence, some partial results are to be memorised in a regular way, and the hearer is conscious that the sentence is ambiguous and that he is resolving the ambiguity. However, a typical ambiguous sentence has one interpretation more easily obtainable than the other ones, and when it is found, the processing of the others is abandoned by the hearer's cognitive processor without making him aware of the fact. It should be noted that in some cases it is the abandoned interpretation which is intended, therefore the participants of the communication act adhere to various rules of "language games", which allow, among others, confirming the proper understanding of the speaker's intentions or quick recovery from misunderstandings.

Most of such rules are still to be found; it might e.g. be interesting to, investigate the polite forms of addressing, asking, etc. from the point of view of their resource consumption, as it may happen that their politeness is due to the optimal processing load distribution in the utterance time span. Independently of this question, the ironical or even offensive character of using over-polite forms may be attributed to the fact that they convey simple meaning in a form requiring heavy resource consumption.

6. RESOURCE CONTROL DEVICES

It is claimed here that every natural language employs various devices for resource control, which allows balancing the demands with the resources available. We distinguish two types of these devices : saliency-related and consumption-related.

The saliency-related devices control the resource consumption only indirectly, by facilitating the retrieval of information from memory. They operate by increasing the saliency of the relevant schemata in the memory, which can be done in two ways : the schemata may be made more central, or they can be assigned more adequate clues (actually, saliency is influenced also by the number of pseudo-physical copies of the schema in the memory, but this factor is rather uncontrollable.

The linguistic constructs used for the purpose range from rhetorical questions to whole passages of texts, recapitulating the information useful for further discussion. Although the control of saliency is rather indirect, it can be quite precise, allowing the hearer not only to find quickly the proper meaning of the utterance, but also to decode allusion and understatement correctly.

Saliency is also and important factor in pronoun resolution, as it is shown by recent research on English (Grosz 1978, Hobbs 1978, Webber 1978). However, it is rather obvious that in languages with grammatical gender the natural way to resolve a pronoun is to find first the relevant nominal phrase using the gender value as a simple means to limit the search (Bien 1976); then either the referent of the noun phrase is taken (a typical pronominal reference) or the phrase is re-evaluated (pronouns of laziness). Therefore we conclude that memory also stores various partial results of the sentence analysis, including some elements of its surface structure. Depending on how carefully the speaker's usage of pronouns takes into account their saliency of the hearer's memory (it is e.g. known that little children are rather bad in this respect), resolving the pronominalisations requires more or less time and resources.

The consumption-related control devices are of very different characters, and are grouped together only for the purpose of the presentation. As one of the most important control device we will interpret the fact that "people tend to express given information, what is already kmown to the listener, before new information, which is not already known. The tendency appears to be universal. Languages overwhelmingly prefer to place definite noun phrases (given information) before indefinite noun phrases (new information). In some instances, the only way to indicate that a noun is definite is to place it before any indefinite noun, or to place it before the verb and to place any indefinite noun after the verb" (Clark and Clark 1977 : 548).

The explanation goes as following : "Given information" is the information which is to be retrieved from memory. The retrieval process takes a relatively important amount of time, because the queries must reach the relevant schemata and then the schemata must be transmitted to the focus of attention. Putting the given information at the beginning of an utterance allows the search to be initiated early enough, so that the results may arrive at the focus area before they are needed. On the other hand, the memory search consumes a considerable amount of resources; because of its low priority it gets only the resources not used by the sorbtion and syntactics. Again, initiating the search as early as possible makes it easier to collect the required amount of resources during the time span between the occurrence of the "given information" and the moment when it is really necessary for further processing of the utterance.

The other aspect of the given before new principle is that the new information is usually put on the end of the utterance. This is quite consistent with our model of the cognitive processor. Because every schema in the focus area, including those representing new information, is spontaneously instantiated, putting new information at the end allows the results of the search for given information to arrive and be integrated into the "new information" schemata. For example, if "the man" in the sentence below

The man bought a book.

means "John Smith", when an attempt is made to instantiate the schema for "book", it is already quite specific because it contains information "bought by John Smith", supplied both by the memory search and the processing of the preceding part of the utterance; in this way wrong instantiations are already prevented. The other factor is that when the utterance is finished, the memory search is probably deprived of most of its resources, which are taken over by the more deterministic processes building the final representation of the utterance meaning during the intersentence gap.

It should be noted that the assumed reluctance to introduce new objects explains very well the essence of the laziness hypothesis of Wilks (1976) and the coherence principle of Bien (1976).

It should also be noted that our given before new principle is actually more general than the Clark and Clark formulation : our claim is that we put on the very beginning those linguistic constructs which require more resource-limited processing. Therefore we are not puzzled by the sentences like

A book was bought by the man.

meaning "the thing bought by the man was a book", which require some amount of inferencing, instead of memory search, to be related properly to its context.

Another important feature of natural language which decreases the processing load consists in mechanisms allowing the creation of some reasonable expectations. We consider two properties discussed in the linguistic literature to be just the examples of such mechanisms : Sgall's systemic ordering of participants (Sgall et al. 1973) and Kuno's empathy rules (1976). Sgall claims that every language has a preferred "neutral" surface order of deep participant (Actor, Object, ..., Adjective of Time, etc.). Although the claim at first glance seems rather strange, it is probably true, and consistent with the ideas advocated here. Similarly, Kuno's examples showing awkwardness of changing the point of view in an utterance, can also be naturally explained in terms of the processor overload caused by the additional task of adapting the processing to a new point of view.

The most interested control devices are related to the syntactics level. They can be divided into global principles, which pose some limitation on possible syntactical constructions, and the local principles, which operate during sentence understanding. To the first type belong some equivalents of Yngve's depth hypothesis, and some resource-allocation rules related directly to specific syntactic constructions. For example, it seems that the main clause top level components always receive more resources than subordinate clauses. It explains why the pronoun in the example (suggested by Richard Hudson)

Mary told him John had won

cannot be coreferential with "John": the main clause pronoun is evaluated immediately, while the schema for "John" becomes available later. Let us discuss now more subtle examples, used already in (Bien 1976):

He went to the pool hall after John left his apartment.

The attempt to resolve "he" starts immediately; it should be remembered that first the proper nominal phrase is looked for. If such a phrase is found and it is not coreferential with "John", then the sentence is properly understood. If the phrase is coreferential with "John", then that means that two independent attempts are made to bring the "John's" schema to the focus of attention. The problem is that both attempts succeed, yielding two copies of the schema searched; this gives the strong impression that, in the best case, two different Johns are talked about.

If there is no suitable antecedent phrase for "he", then there is another kind of problem, because "John" is processed too late to contribute the relevant information to the top level of the meaning structure of the main clause; therefore such a usage of the sentence is incorrect. On the other hand, we can use the cataphoric reference in the sentence

After he left the apartment, John went to the pool hall.

because "he" belonging to a subordinate clause waits for processing long enough,due to the small amount of resources allocated for subordinate clause processing.

The syntactic structure of an utterance may also influence the semantic processing in a much simpler way, e.g. it may suggest the proper environment for the interpretation of noun phrases embedded in some intensional context (Bien 1976). However, the syntactic control device of crucial importance for Slavic languages relies on the following assumptions. One of the tasks of the syntactics level is to generate predictions about the part of the utterance yet to be analysed. Obviously, generating the prediction consumes some resources; the number of predictions varies during the utterance analysis (if it is too big, the syntactics become blocked by the resource allocation rule).

Although Marcus-style parsers for Slavic languages are yet to be written, it may be expected that when the analysis arrives at the element, which according to the systemic ordering would be expected later, no reasonable predictions can be generated. The resources released by the syntactics are then used for memory search in the case of e.g. definite reading of a nominal phrase, or for some inferencing, e.g. in the case of emphasis. In this way we can supply an explanation for the fact noticed by Sgall that the element moved forward relative to its default position usually receives the definite reading.

As a separate category of resource control we consider various phenomena often interpreted as "redundancy". It was already noted by Wilks (1975) that so called redundancy (Joos 1972, Scriven 1972) is useful for properly resolving some kinds of ambiguities. We agree with this observation, but we concieve the essence of the phenomenons slightly differently. For example, Joos' semantic axiom No. 1 that "the best meaning is the least meaning" we assume to be just a side-effect of the full capacity principle: the speaker should supply the new information in relatively small portions if he wants to be understood easily.

A good way to facilitate the hearer's task is to allocate a longer time span by inserting into the utterance some words requiring little processing, e.g. "redundant" demonstratives. This method is sometimes used in Slavic languages (Szwedek 1976) to obtain the effect of definiteness; in languages like English, this was the main method, which resulted finally in converting the demonstratives into the definite article. The indefinite article and its rarely used Slavic equivalents like the numeral "one" resemble rather the category described earlier, the difference consisting in the fact that the hints about low resource allocation (to disallow memory search) come from the lexical meaning instead of from specific syntactic clues.

The above list of resource control devices is not intended to be complete. On the contrary, one should expect new control devices to be discovered in the future. For example, the recent findings of Rachel Reichman (to appear) concerning the differences in usage of "the", "this" and "that" and simple present versus present continuous tenses may be interpreted as new types of control information. As far

as spoken language is concerned, it is possible that the elements stressed are easier to analyse by the sorbtion level because of their acoustic properties, so the surplus of the processing capacity is allocated to the higher levels yielding e.g. the effect of emphasis.

The interaction of various control devices in the same natural language is yet to be investigated. One can expect some kind of specialisation, e.g. it might be the case that in German the articles control mainly the memory search, while the relatively free word order is used to control spontaneous inferences.

7. CONCLUSION

Although not yet substantiated satisfactorily, the resource control hypothesis opens new perspectives, allowing us to view a variety of seemingly unrelated linguistic phenomena as the sophisticated interactions of a few basic factors. In consequence, a general, intuitively appealing, framework for natural language processing is proposed, within which it is hoped to accomodate in the future all well established results of language understanding research.

8. REFERENCES

(Allen, Perrault 1978)
Allen,J.F. and Perrault, C.R. Participating in Dialogues Understanding via Plan Deduction. AI-MEMO 78-4, Department of computer science, University of Toronto, July 1978. Also in: Proceedings, Second National Conference, Canadian Society for Computational Studies of Intelligence, Toronto 1978.

(Ashby 1952)
Ashby, W. Ross. Design for a Brain. London, Chapman & Hall, 1960 (second revised edition; first edition, 1952).

(Bien 1975)
Bien, J.S. Toward a Multiple Environments Model of Natural Language. Advance papers of the Fourth international joint conference on artificial intelligence, Tbilisi, 1975, pp. 379-382.

(Bien 1976)
Bien, J.S. Multiple Environments Approach to Natural Language. American Journal of Computational Linguistics, microfocihe 54, 1976.

(Bien 1976a)
Bien, J.S. Computational Explanation of Intensionality. Reprint no. 41 of the International Conference on Computational Linguistics, ottawa, 1976.

(Bien 1977)
Bien, J.S. Ph.D. Thesis in Polish, Institute of Informatics, University of Warsaw, 1977.

(Bobrow, Norman 1975)
Bobrow D.G., Norman D.A. Some principles of memory schemata. In: D.G. Bobrow and A.M. Collins (Eds.), Representation and understanding: Studies in cognitive science. New York: Academic Press 1975.

(Clark, Clark 1977)
Herbert H. Clark, Eve V. Clark. Psychology and Language. An Introduction to Psycholinguistics. New York, Harcourt Brace Jovanovich, 1977.

(Cohen 1978)
Cohen, P.R. On Knowing what to Say : Planning Speech Acts. Technical report no. 118, Department of computer science, University of Toronto, January 1978.

(Grosz 1978)
Barbara J. Grosz, Discourse knowledge. In D.E. Walker (Ed.), Understanding Spoken Language, New York: North Holland 1978.

(Hewitt, Bishop, Steiger 1973)
Carl Hewitt, Peter Bishop, Richard Steiger. A Universal Modular ACTOR Formalism for Artificial Intelligence. IJCAI-73, pp. 235-245.

(Hobbs 1978)
Jerry R. Hobbs, Coherence and Conference. SRI International Artificial Intelligence Center Technical Note 168. Also in Cognitive Science, Vol. 3, No. 1.

(Joos 1972)
Joos M. Semantic Axioms No. 1 Language. 1972, 193-211.

(Kaplan 1973)
Ronald Kaplan. A Multi-Processing Approach to Natural Language. National Computer Conference 1973, pp. 435-440.

(Kowalski 1979)
Robert A. Kowalski. Logic for problem solving, New York: North Holland 1979.

(Kuno 1976)
Kuno S. Subject, Theme, and the Speaker's Empathy. A Reexamination of Relativization Phenomena. In Li C.N. (Ed.), Subject and Topic, Academic Press, pp. 417-444.

(Lynn Webber 1978)
Bonnie Lynn Webber. A formal approach to discourse anaphora. BBN Report No. 3761.

(Miller, Johnson-Laird 1976)
Miller, G.A. and Johnson-Laird, P.N. Languages and Perception. Cambridge, Mass., Harvard University Press, 1976.

(Norman, Bobrow 1975)
D.A. Norman, D.G. Bobrow. On Data-Limited and Resource-Limited Processes. Cognitive Psychology 7, pp. 44-64, 1975.

(Norman, Bobrow 1976)
D.A. Norman, D.G. Bobrow. on the Role of Active Memory Processes in Perception and Cognition. C.N. Cafer (ed.). The Structure of Human Memory. San Francisco, Freeman.

(Norman, Bobrow 1979)
D.A. Norman, D.G. Bobrow. Descriptions: An Intermediate Stage in Memory Retrieval. Cognitive Psychology 11, 107-123.

(Perrault, Allen, Cohen 1978)
Perrault, C.R., Allen, J.E. and Cohen, P.R. Speech Acts as a Basis for Understanding Dialogue Coherence. AI-MEMO 78-5, Department of computer science, University of Toronto, July 1978.
Proceedings from Theoretical Issues in Natural Language Processing II. University of Illinois at Champaign-Urbana, 1978.

(Reichman, to appear)
Reichman R., Talmudic Exegesis of Natural Conversations. To appear as ISSCO working paper.

(Scriven 1972)
Scriven M. The concept of comprehension. In Caroll and Freedle (Eds.), Language Comprehension, Washington, D.C., 1972.

(Sgall et al. 1973)
Sgall P., Hajicova, E., Benesova, E. Topic, Focus and Generative Semantics. Kronberg, Taunus.

(Shipman, Marcus 1979)
D.W. Shipman, M.P. Marcus. Towards Minimal Data Structures for Deterministic Parsing. IJCAI-79, pp. 815-817.

(Smith, Rawson 1976)
R. Smith, F. Rawson. A Multi-processing Model for Natural Language Understanding. COLINY-1976.

(Szwedek 1970)
Szwedek A. Word Order, Sentence Stress and Reference in English and Polish. Edmonton, 1976.

(Wilks 1975)
Yorick Wilks. An Intelligent Analyser and Understander of English. Communications of the ACM, Vol. 8, No. 5, pp. 264-274.

(Wilks 1975a)
Yorick Wilks. A Preferential, Pattern-Seeking, Semantics for Natural Language Inference. Artificial Intelligence 6 (1975), pp. 53-74.

(Wilks, Bien 1979)
Wilks, Y. and Bien, J.S. Speech Acts and Multiple Environments. proceedings of the Sixth IJCAI, Tokyo 1979, pp. 968-970.

(Ziff 1972)
Ziff, P. What is Said. In D. Davidson, G. Harman (eds.). Semantics of Natural Language, pp. 709-721.