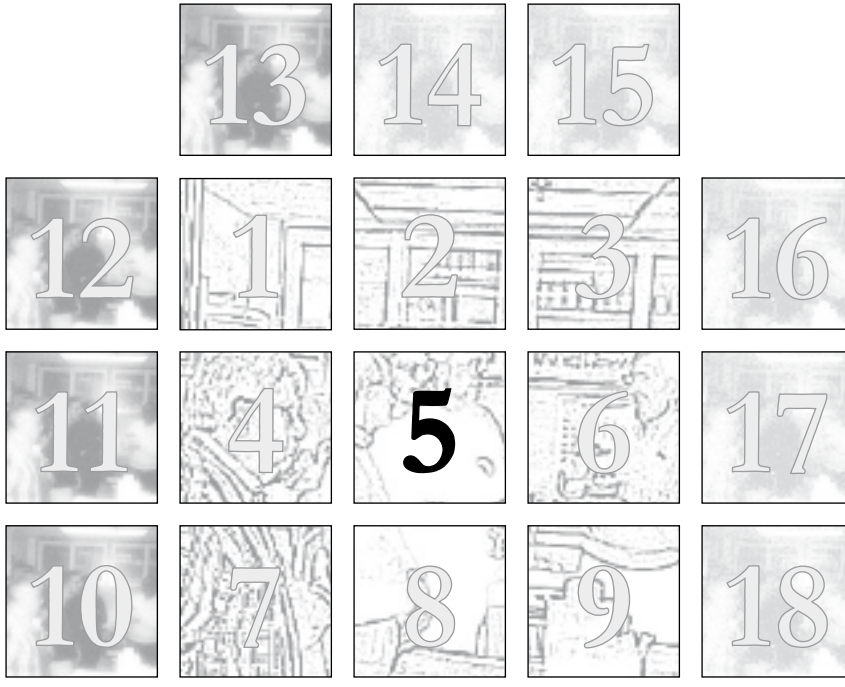
5 A Walk Through the Forest

Paul Weston



I would like to thank the conference committee, the sponsors, and especially Dr. Müller for extending the invitation for me to speak at the 2003 Von Foerster conference and to contribute to this volume. It was a pleasure to see and hear once again those of the original BCL who were still with us, and the many younger people who share the legacy and are continuing the course that was begun half a century ago.

The title of my original talk, "A Walk Through the Forest", was chosen for several reasons:

First: metaphorically, it was for the younger among us, including me, a journey in unexplored territory. There were highly detailed "trees" on all sides of us, bidding for our attention and tempting us to become lost in over-focused detail. Through Heinz's leadership we were joined by a steady stream of men of vision, Ross Ashby, Gordon Pask, Gotthard Günther, Humberto Maturana, to name a few from the early days, who kept us aware that we were in a vast and living forest.

As a sidelight, it is a pleasant thought to realize that Heinz was our true and our metaphorical "Foerster", or in English, forest warden, who held the compass and the map and kept our spirit of adventure alive.

Finally, I simply enjoy the image of a forest, having lived my early years in the middle of a real one. That forest provided some necessities of life; it was a playground, a school, a means of travel, and held places of true beauty. The real BCL did many of those things.

In Fall of 1958, I believe, a young physics graduate student at the University of Illinois, who had been excitedly reading such things as Norbert Wiener's little book, "Cybernetics", attended a seminar on that very subject given by a professor in electrical engineering. The speaker was short and somewhat balding, spoke with a definite German accent, had an engaging personality, and lectured very well.

As it happened, the young physics student was then finding his readings in Cybernetics distinctly more stimulating than his experimental work on short-lived excited states of atomic nuclei. So he found himself some time later in the office of the professor who had given the seminar, exploring the possibility of becoming involved in his (then) small research group. That involvement materialized. In the fall of 1959, the young physics student became a young electrical engineering student in what was soon to be known as BCL.

Our first encounters in the "forest" of Cybernetics were neural networks, both artificial and living ones. We wanted to see how far we could go toward making machines that could behave like living organisms. New and revealing biological research was being done at the time on the nervous systems of animals, including the classic study by Lettvin, Maturana, McCulloch and Pitts.

"What the Frog's Eye Tells the Frog's Brain". Although the attempt to build neural networks was not ultimately successful, it seemed natural to imitate nature by designing systems with neuron-like elements. Because the available animal research results were mostly in the area of sensory systems, the general problem of pattern recognition came to the forefront. This includes the broad field of mechanisms through which a machine or an organism acquires the ability to respond to invariant properties of varying stimuli, particularly those properties which are important to the self-maintenance of the organism, or the intended function of the machine.

Under less competent leadership this could have been just an engineering exercise doomed to failure. In BCL it was seen in the broader cybernetic context of adaptation and self-reference. (Gordon Pask joined us during this period). For practical reasons we soon abandoned attempts to make machines with simulated neurons, although a small adaptive machine on those principles was demonstrated by Murray Babcock (who, sadly, passed away in 2000). The fundamental problems of reference and self-reference and self-organization remained very much in our attention.

Of course, *theoretical* work on neural nets continued at BCL and elsewhere for some time. At BCL, Heinz published several stimulating theoretical papers, a doctoral thesis was generated by student Ronald Swallow, and very importantly, enduring contacts were established with Humberto Maturana and Ricardo Uribe.

Since it had a bearing on the direction of practical, and secondarily of theoretical work in the field, let me digress briefly to mention the nature of the problems we had in realizing actual machines.

To simulate a neural network, for instance, means are needed for representing the transmission strengths of a large number of synapses; in other words a lot of numbers have to be stored, retrieved, and modified. This is trivially easy with present storage technology, but all but impossible over forty years ago at the scale we needed.

The largest computer at the Urbana campus in the mid 50's had a tiny fraction of the power of a present-day handheld, took up the space of a large room and used 20 large CRT assemblies to store its 10KB of RAM which held both program and data. Only the very earliest commercial transistors were entering the market toward the end of the decade, and they were expensive and underperforming. Kilby's revolutionary integrated circuits had not appeared. Magnetic core storage was very costly and somewhat unreliable; similarly for rotating magnetic storage systems.

Some partially-electronic schemes that now seem bizarre were tried, including heat storage in liquid and solid media and reversible electro-plating processes;

even a shape-remembering metal alloy was examined. Gordon Pask did have limited success in his adaptive teaching machines with metal filaments crystallized and redisolved in a tightly controlled electrolyte medium. But it has only been in the last twenty years or so that computers have been able to do the rapid, large-scale computations really required. (And in that period neural networks were rediscovered by a younger generation.)

5.1 First Project: The Numarete

Shortly after joining the BCL group, I, the erstwhile young physics student, built a machine without simulated neurons which could recognize an abstract property common to a large set of distinct stimulus patterns. It contained an iterated network of identical devices which gave the outward appearance of a possible neural network model, and that has led to various inaccurate descriptions of the device. The device was dubbed the "Numarete" which is a simple anagram of "numerate", which in turn is a synonym of "enumerate". Its function was to count the number of distinct objects of any shape and position that might be presented to its photocell "retina". The story of its birth and death are told in the Festschrift, and won't be repeated here. What will be done is to show what is inside, revealing how little intelligence it really had. However, many uninitiated persons who tried and failed at the time to "trick" it into error by, e.g., inserting objects within holes in larger objects were ready to believe that it was intelligent.

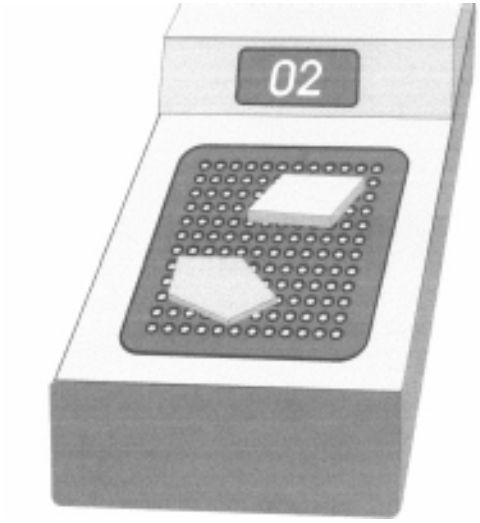
Figure 1 represents the appearance of the Numarete when it flew to New York City. No actual photographs of that configuration remain. The "retina" was a twenty by twenty array of four hundred photocells. (This sketch shows only 144). Note that objects block the ambient light from reaching the photocells beneath them. The numerical readout had actual well-formed numerals as shown here, using a now-obsolete type of gas-discharge display tube called the NIXIE. That was before the age of seven- or fourteen-segment solid-state numerical displays.

Control signals for the display tubes came from an electronic counting circuit, which by that time was a standard item, available in many forms. The overall action of the counting circuit and the display was to advance the displayed value by one, for each brief voltage pulse applied at the counter circuit input.

Each photocell in the Numarete retina is directly connected to a small electronic computing cell whose functional structure is shown in block diagram form in Figure 2. A one-dimensional array is shown, since that is enough to explain

the mechanism. The actual device differed only in having four rather than two pairs of lateral connections, two for the nearest neighbors along rows, and two for the nearest neighbors along columns.

FIGURE 1 **The Numarete**



The computing cell has only two possible states, call them ON and OFF. At the start of a counting cycle all the cells are reset to the OFF state. The *only* two means of turning a cell OFF are this initial reset action and having light shine on its photocell. A cell is disabled and *cannot* be turned ON if its photocell is in the light, i.e. outside the boundaries of any object.

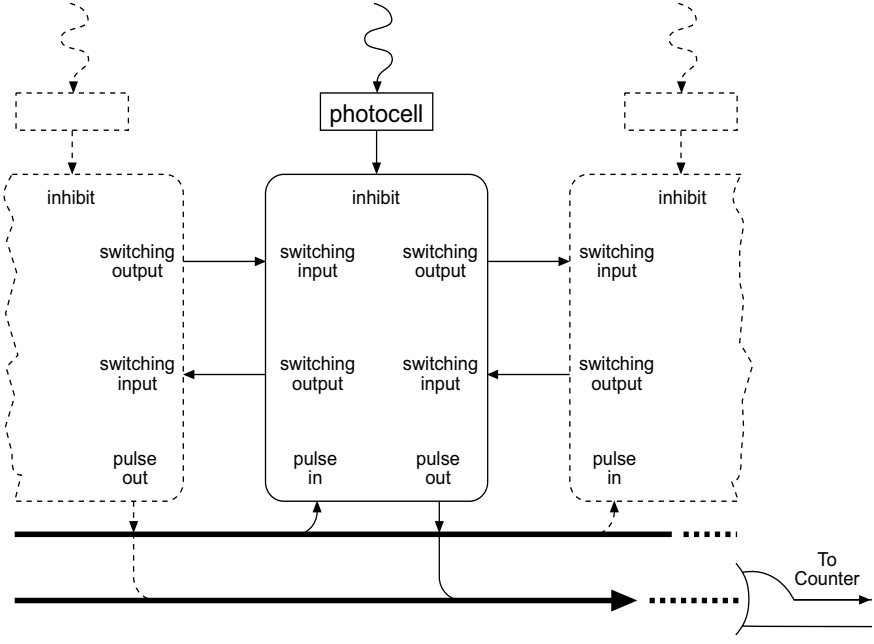
Any cell that is under an object, and not already ON, can be turned ON by applying a brief signal to its input labeled "trigger pulse". It will then cause all of its nearest neighbors which are in shadow to turn ON, and they turn ON their nearest neighbors if they are in shadow, and so on until all the cells shaded by the same object are ON. This is the function of the lateral connections labeled "switching input" and "switching output" in Figure 2. The signals propagate from cell to cell so rapidly that the entire group associated with any one object appear to switch ON at once.

We now see that all the cells under a given object will turn ON together if *any one* of them is turned ON by some means. Keeping in mind that the counting cycle is initiated by forcing all cells OFF, the question becomes how that first cell is turned ON, and how the actual counting occurs.

The answer to the first part of the question revolves around the cell connection labeled "trigger input" in Figure 2. In the Numarete control circuitry a means

is provided for sending signals to the trigger inputs, one at a time, on each of the cells in the network, in a simple row-by-row scanning pattern. If the cell being scanned is not under an object it cannot turn ON, and nothing happens. If the cell is under an object it may already be ON because another cell under the same object was reached earlier in the scanning sequence, and once again, nothing happens.

FIGURE 2 **Numarete Computing Cell**



If the cell that is receiving a trigger signal is the first one under a particular object to be reached in the scan, then all cells under that object will turn ON together at that time, and all will generate signals at their own "output pulse" terminals (see Figure 2). All of these signals reach the input of the logical OR circuit together (the circuit is shown in the lower right-hand corner of Figure 2), and it, in turn, generates a single output pulse which is sent to the input of the electronic counting circuit (not shown).

The defining property of the logical OR circuit is that its output signal is generated if an input signal appears at any *one or more* of its input connections, so that, while many cells are simultaneously sending their output signal pulses to the input of the OR circuit, it produces only a single output pulse, which advances the electronic counter by one unit. Since only one cell at a time receives a trigger input, and all cells under a given object respond instantly

for practical purposes as soon as the first one is triggered, and thereafter are unresponsive to trigger signals, the counter receives exactly one pulse during the counting cycle for each separate object lying on its "retina".
It is clear that the machine described above is neither modeled upon biological neural networks nor in any important sense intelligent, though it does contain an array of identical computing cells, which might suggest the appearance of a neural network, and it does report the same number of objects as a human observer in a fraction of the time. It is, perhaps, an example of the general rule that technology can duplicate aspects of biological performance but almost never by using exactly corresponding mechanisms.

5.2 Beyond Pattern Recognition

In the immediately following years, BCL students continued to generate theses based upon realizable hardware and software, focusing on sensory mechanisms and recognition of patterns both auditory and visual. But as I saw it, they all had a fatal flaw; no account was being made of the clear fact that humans do not see or hear the same thing every time a particular objective stimulus is received. No matter how perfect a machine might be in detecting subtle properties of sonic or visual patterns, it could never *respond* in anything like a human way. The subject her- or himself plays an active role in the perceptual process, contributing materially to the result. This point probably does not need to be further belabored before this particular audience.
If anyone does have doubts he should try listening to a short endless loop of random noise. After a while, only on the order of a minute, one hears a breathy, hollow voice clearly saying a *word or phrase*. After continued listening, the utterance may change but it remains intelligible.
Forty years ago it seemed to me that I should leap-frog over the pattern recognizers and look into language, as far as I was able, since it was the meanings of things to people which appeared actually to govern the perceptual process, possibly the very same kind of meaning that is expressed through language.
During the time-consuming process of absorbing background literature, from Carnap to Wittgenstein, Chomsky, Sapir, and the papers of the then current research by others who had taken up the task, some early topics came up which could be looked into without reference to the logico-philosophical side of the issue. In particular, there were two statistical studies: one dealing with word frequencies in written text, and another with relations among word definitions in the vocabularies of several languages.

5.3 Statistical Regularities of Language: Zipf

The first was a review of Zipf's law, formulated by George Kingsley Zipf, who was active in the first half of the last century. The law describes an unexpected statistical regularity in word frequencies, which appears across different languages, authors, topics, etc. If word frequencies are counted in a given body of text, and ranks are assigned in descending order of frequency, i.e., the most frequent is assigned rank 1, then, over most of the words, rank and frequency are inversely proportional. On a log-log plot the slope is often a bit steeper than minus one, but seldom more than 1.4. Zipf seems to have assumed that it was ideally exactly minus one.

It was not long before it became known that this particular statistical behavior of word frequencies could be duplicated in a trivial way. That is, the well-known monkey at the typewriter produces text with word frequencies obeying Zipf's law from the beginning, long before he creates a Shakespearean sonnet. While some took this discovery as a simple 'explanation' of Zipf's law, Zipf himself held otherwise, though he was never able finally to resolve the issue.

I sided with Zipf. Demonstrating a mechanism which duplicates only one aspect of a given behavior is hardly an explanation of that behavior. Consider the Numarete, described above. It flawlessly reports the number of objects that it 'sees,' yet it uses a mechanism which has almost no similarity to the action of biological neural networks, and can do nothing with the numbers that it reports. It clearly does not explain any aspect of human vision or the process of counting, as done by human beings.

To illustrate one significant difference between random text and that written by human authors, the monkey's random strings of letters, his "words", are separated from each other by the typing of space characters, which in the case of the monkey is presumed to occur with constant probability. This causes longer words to be less likely, with their probability decreasing exponentially with increasing word length, and this, in turn, leads to the rank-frequency relationship described in Zipf's law. Human text does not show any such simple relationship between word length and frequency.

It does not appear that randomness can explain the adherence to Zipf's law of the highly structured text written by people. The monkey, typing at an expert stenographer's rate since the beginning of the universe, would have had almost no chance of putting together a single paragraph of intelligible material by the present day. Writing intelligible text is far from a random process. Words are selected and combined into sentences, often with the utmost care. They are chosen to convey particular ideas and must adhere to complex syntactic rules. In none of this is the writer of the text consciously

concerned with word frequency. Can *any* explanation be found? In 1951, the mathematician Benoit Mandelbrot published a short paper, using the then relatively new mathematical theory of information, showing that the statistical properties of Zipf's Law may contribute directly to *efficiency* of communication. He introduced the notion of the "cost" of decoding (i.e., in this context, reading) words, and showed that if that cost were related to word frequency in a particular, plausible, way, then the observed rank-frequency relationship would achieve the least possible cost to the decoder (i.e., the reader). That is, messages *designed* to convey information will do so most efficiently by adopting the statistical structure of purely random strings of symbols (recalling the monkey at the typewriter).

The question becomes: how must this cost be related to word frequency, and can any such variable be identified in the behavior of actual human readers? Regarding the first part, Mandelbrot demonstrated that it is only necessary for the cost of reading a word to vary linearly with the logarithm of the word's frequency in the given body of text. In information theoretic terms, this means that the cost of decoding should be proportional to information content. Regarding the behavior of human readers, findings had been reported by Howes and Solomon at the time of Mandelbrot's paper which showed that the time *required* to read a word is independent of the word's length but does vary linearly with the logarithm of frequency, i.e. with information value, as required of the cost variable. Referring these facts back to Mandelbrot's efficiency argument, with time playing the role of cost, we can conclude that the peculiar rank-frequency statistical structure of language enables human readers to assimilate the greatest amount of information per unit of time.

What we cannot conclude from this is what actually goes on in people's heads when they read or write in natural language (as opposed to artificial languages, such as programming "languages"). On the other hand, the above does make clear that any particular word has no *intrinsic* quantitative information value, in the information-theoretic sense. It may have almost any relative frequency, or not appear at all, in different texts dealing with different topics and by different authors. We tend to believe that meanings of words are constant properties or at most slowly changing, yet we find that the words convey no fixed amount of information. Apparently meaning and information value are not as intimately related as they might seem.

5.4 Human Language from the BCL-Perspective

To set the stage, the following paragraphs paraphrase a 1963 paper of mine entitled "Machine Use of 'Natural' Language". They are offered to shed light on our thinking at that time in BCL about problems relating to human language.

Our central idea was that language functions by directing the reader or listener to construct a mental representation, and that this representation is of the very same sort that he might construct in the direct perception of objects and events in the world (external and/or internal to the person). In other words, language is used to 'externalize' mental representations, *thoughts*, if you will, and can only *indirectly* refer to external realities through our internal representations of them; it can succeed in the latter only to the extent that these representations have properties that are shared between speakers and listeners or writers and readers, and, most importantly, that these properties are also mirrored in the structure of their shared language. Indeed, humans are able to envisage and mentally manipulate an unlimited variety of things which do not, and often cannot, exist in external reality and one would expect to find reflections of this additional richness in their language.

The radical sort of philosophical relativism implied by the position sketched above was not in the main stream at the middle of the twentieth century, and others more qualified than I have carried the philosophical debate up to the present day. This concept of the function of language emerged for us, nonetheless, as the most reasonable one, growing out of our interest in the simulation of the processes underlying perception. It served to motivate the extension of our interest beyond pattern recognition principles into the exploration of how information is conveyed in the syntactic and semantic structure of language.

While the rules of grammar in any particular natural language are largely intuitive to native speakers, they are quite complex, taking considerable time, for instance, to be fully assimilated by a person learning the language. At about the time of BCL's founding, Noam Chomsky was publishing his ground-breaking work on transformational grammar, which for the first time succeeded in capturing the subtlety of grammatical rules in an elegant formal scheme. His theory begins from a set of kernel sentences generated by a small set of phrase structure rules and taking the form of simple declarative sentences in present tense. These are greatly augmented by a set of transformation rules by which expressions of tense, active vs. passive voice, interrogative or imperative mood, etc. can be correctly generated from the kernel sentences. While Chomsky wisely eschewed any explicit involvement with semantics in

formulating his theory, it is clear that his transformation rules implement the introduction of information that is not contained in the basic kernel sentence and that is essential to the reader or listener, regarding time reference or assumed relationship of speaker to listener, etc. These information-bearing functions of syntactic rules, which operate independently of specific referential content of the sentence, provide useful glimpses of what must be the *framework* of an internal representation, which we were willing to assume is also the framework in which the world is directly perceived.

Even the glimpses of that framework that can be found in grammatical transformation rules verify the intuitive impression that the task of designing a faithful simulation of human understanding of natural language is truly monumental, far beyond the capacity of BCL then, or any other organization, up to the present time. Such a simulation would need to encompass the span of human emotion and motivation, and be able to formulate abstract 'thoughts' and understand metaphors, etc., etc. While the term 'artificial intelligence' has now entered the general lexicon, its practitioners are still obliged, as were those of forty years ago, to work within narrow domains in which the relevant logical and semantic relationships have been worked out ahead of time and reduced to computer programs and data structures.

5.5 The Noun-Chain Project

The context of this early project came from considering the following observation in light of the previous discussion: no one can possibly be aware at once of everything in the world, or even in his own small local region of it.

Perception inherently involves a focusing of attention, a partly conscious and partly unconscious restricting of the sources of current or remembered sensory data *and* of the sorts of interpretation to be applied to them. One directs one's gaze in a particular direction, touches a particular object, listens for or to a particular source of sound, feels the soreness in a particular muscle, sees in his/her 'mind's eye' a particular place previously visited, and so on. The mere act of focusing of attention only *enables*, but does not *accomplish*, the gaining of new information. New information is internalized by adding to the existing internal representation, or 'context', that has been formed from events up to the present time (at least the fairly recent events; this is probably what we think of as short-term memory). At very least the process involves the specific content of the present context, general knowledge of the world, and complex innate sensory processes which extract perceptually useful data from the field of attention, such as binocular vision.

Consider this possible scenario; a sudden sound attracts my attention, and I quickly realize that it is made by my dog, who is lying on the floor to my right (my representations of my location in the room, the other objects around me, the location of the dog, etc. all being parts of the current context). In a reflexive response to the sudden sound, I have shifted attention to it; innate mechanisms of binaural hearing, which process temporal and spectral differences between what reaches my two ears, and the timbre of the sound have allowed me to attribute it to the dog, the perceived direction being consistent with the context. Note that seeing this consistency requires general knowledge of three-dimensional space, and of dogs. Based upon the newly-extended context, I may then turn my attention to a window, a door, a food dish, and so on, to see what may have elicited the dog's bark originally.

There are, then, two inseparable and complementary basic parts of the perceptual process. First, there is the focusing of attention, which operates internally both to select elements from the current context and to select sensory channels. Second, the active internal construction of a representation to fit the selected sensory events into an extension of the context, wherein new information is actually gained.

In our view, wherein language is based upon the mechanisms of perception, the logical structure of language into sentences containing subjects and predicates, and the nonsensical nature of incomplete sentences, directly mirrors the dichotomy of attention focus and information collection, and their inseparability. That is, some parts and/or structural features of the sentence must function to select some part of the *current* context (although the focusing of sensory channels is not directly involved), and others must convey an extension to that context, that is to be constructed.

Looking at simple present-tense declarative sentences (Chomsky's kernel, from which all others can be derived by transformation), our intuitive reactions to sentence fragments reveal that the subject of the sentence, a noun phrase, carries the selective function and the predicate, a verb phrase, carries the constructive function, i.e., the 'new' information. While a noun phrase serving as the subject of a sentence may contain elaborate internal structure, its combined function is only to refer to the structure of the current context (or, less frequently, to *establish* a new sub context; this happens in the situation of starting to read a story, where the reader is aware of the book itself as part of the current context and of the fact that he is reading it, but must 'open a new window', speaking in present-day computer terms, to represent the subject matter of the story. Think of "Once upon a time, in a galaxy far, far, away ..".). The verb phrase, similarly, regardless of internal complexity, supports the construction of extensions to the preselected portion of the

context. The fact that the verb inherently carries a time reference indicates the fundamental nature of time in the construction of a mental representation. Within the domain of kernel sentences the time reference is unchanging, and therefore can be ignored. When that is done, there is no further conceptual distinction between noun phrase and verb phrase. Both must be able to refer to arbitrary linguistically expressible elements or structures within an internally represented context, and the distinction between their fundamental roles in the sentence interpretation is the sole (but very important) reason for their distinction as separate parts of sentences.

Were this a treatise on what I believe is currently called ' cognitive grammar', there would be more to say than appears above, particularly regarding the open-ended recursive internal structure of noun and verb phrases and the various parts of speech involved in this structure. The noun-chain study, however, was undertaken as a preliminary exploration, and was not particularly involved with the syntactical domain. The thinking went as follows: if the function of language is to convey mental representations, and words and syntactic rules are the elements of language, then any exploration of the relationships among words might possibly reveal corresponding relationships within mental representations, and that could lead toward more powerful simulations of human language behavior. A dictionary, it was thought, might be an interesting ready-made laboratory for such an exploration, since it consists of a large corpus of words intentionally organized to reveal word-to-word relationships. Among these, the noun was expected, in our way of thinking, to be the least encumbered with syntactically-related overhead.

The plan of action was grounded in the assumption that the definition of each word would be based upon one with more general meaning, whose definition would lead in turn to another still more general one, generating a chain of definitions which would have to terminate at some term so general that it could only be defined by using synonyms of itself. The result of exploring these chains would be a set of 'family trees' of related words, each tree rooted on one of the most-general terms. The distance from the root of the tree would be a rough indicator of the relative specificity of the word's meaning, and the set of most-general terms might yield some insight into the nature of a mental representation. We were setting off on a stroll through a 'forest' of words.

It was immediately discovered with a little surprise that lexicographers do not adhere in general to the neat hierarchical organization described above. Adjectives and adverbs are commonly defined upon either verb or prepositional phrases in which nouns carry most of the content, a pattern which completely breaks the hierarchical scheme. Verb definitions are indeed based upon other

verbs, but usually ones with very general meanings, combined with highly specific adverbial phrases which carry most of the content, so that all their chains appear to be only one or two links in length. Noun definitions, as we had thought, do fit the hierarchical pattern well and are more numerous than those for all other parts of speech combined.

FIGURE 3 **An English Noun-Chain**

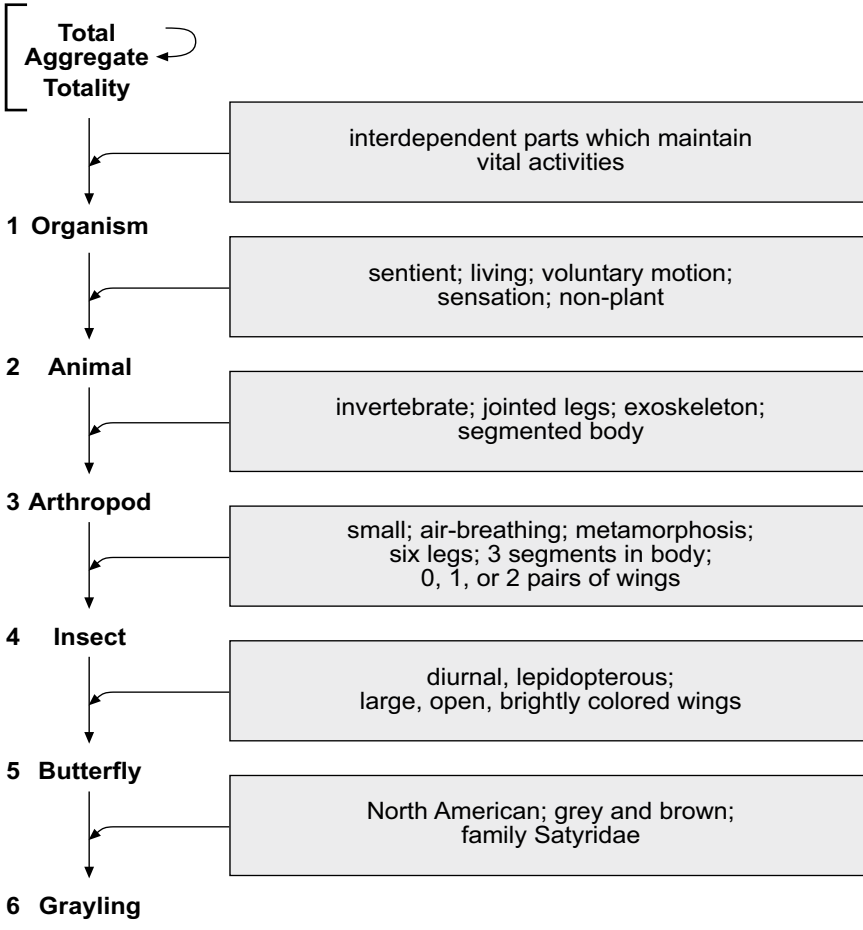
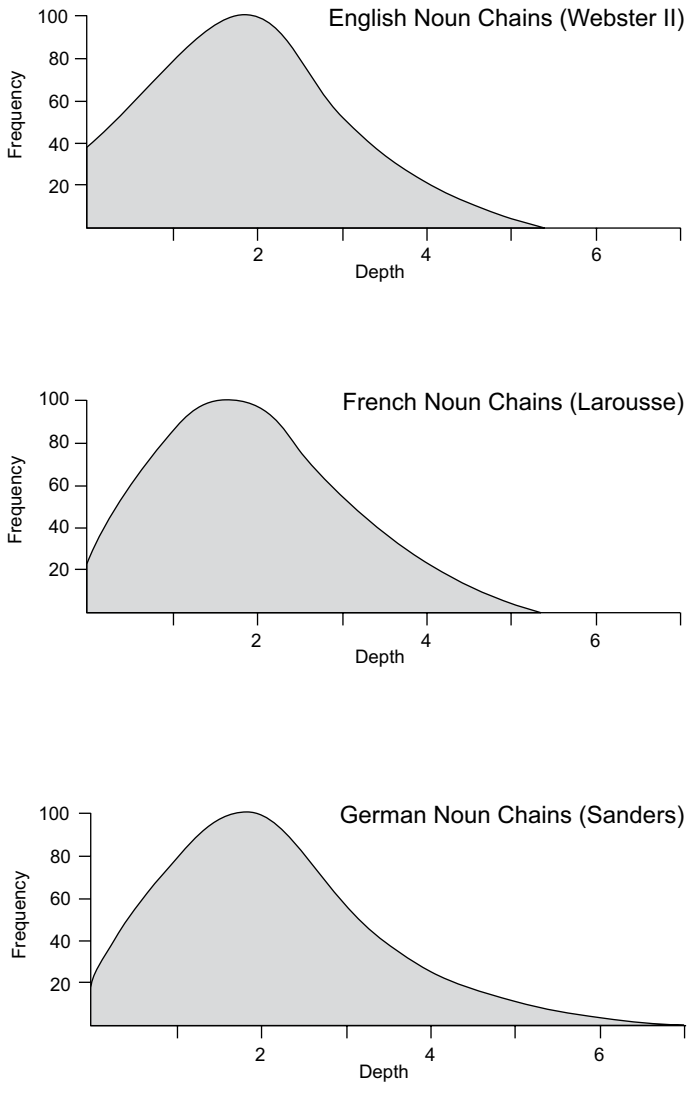


Figure 3, by way of illustration, shows the structure of a fairly long noun definition chain in English. The dictionary entries in the chain are shown beside numbers indicating their distance from the root, which is shown at the top of the figure, consisting of a cluster of three terms. In the dictionary from which this chain was taken, those three root terms, "total", "totality"

and "aggregate" are all defined in terms of each other, that is, in the context of this definition chain the three are synonymous. The boxes in the figure list the properties which differentiate the terms at successive levels in the chain.

FIGURE 4 **Noun-Chain Frequencies vs. Chain Length**



In all, three hundred chains were examined in each of three languages, English, French and German, for a total of nine hundred. The starting points for the chains were determined randomly, by finding the first word whose principal definition was as a noun (excluding gerunds) on each of three hundred equally-spaced pages in the dictionary, and the work was done by assistants who were fluent in each language.

The results did not reveal all the things that were hoped for. In particular, there were far too many separate trees to allow for their root terms to be easily interpreted as representational dimensions or building blocks, but on the other hand a completely unexpected regularity appeared in the statistics of the noun-trees across the three languages.

For each of the three languages, Figure 4 shows the frequencies of occurrence of the various chain lengths within the set of three hundred sampled chains. (Bezier lines have been constructed only as an aid to better visualize the shapes of the distributions.)

While there are, indeed, significant differences among these three frequency distributions, the more striking and unexpected result is their very high degree of similarity. In each case chains of length one and two are more numerous than all other lengths, length two has the highest frequency, and there is a rapid fall-off to negligible frequency around length six, giving all three curves highly similar shapes. The most notable differences are that the tail of the German distribution is definitely more extended (i.e., there are relatively more chains of length greater than two) and there are relatively more root terms (chains of length zero) in the English sample.

With the anticipated results failing to materialize, the question then became one of finding some mechanism that could explain the presence of this uniform statistical structure in the vocabularies of the three languages, and presumably others as well. Just as in the case of the Zipf statistics, where writing is done with no awareness of the resulting frequency statistics, the teams of lexicographers who constructed the dictionaries were simply reporting on the common knowledge of the speakers of the languages and could not have purposely imposed the uniformity revealed here.

A logical first step in the quest for a mechanism was to try to find a known statistical distribution function which could be fitted to the chain data, since such functions lead directly to the kinds of processes which give rise to them. The curves shown in Figure 4 are highly skewed, which narrows the field of usable distribution functions, but the 'tails' fall toward zero at a faster-than-exponential rate, and the relatively few skewed statistical functions which arise in engineering practice, such as the Poisson distribution or the exponential distribution, could not be fitted to the sharply descending tails of our curves.

In a wider search, an interesting alternative from the field of Botany was brought to my attention, the Willis distribution. J. C. Willis, working in the early twentieth century, studied the world-wide distribution of plant species in support of his thesis, for which he made a fairly convincing case, that new species are not the result of many incremental evolutionary changes, but emerge at long intervals as single larger mutations. One result of his work was a family of skewed distribution functions describing the expected numbers of species over time and land area, the details having been worked out by his associate, G. U. Yule. I believe it was Heinz who put me in touch with this material, delighting in the intriguing parallels that might be drawn between the spread of word 'species' and plant species.

The Willis distributions, however intriguing their prospects for further theorizing, proved entirely unsuitable for our data. Since no other viable candidates were unearthed and little progress was made in extracting a mathematical description of theoretical interest directly from the data, the matter was laid to rest forty years ago, only to resurface in the preparations for the 2003 Von Foerster conference. I found then that I still had some records on that project, with its still unsolved puzzle. Running the data through various hoops in a modern spreadsheet program (which was unknown at the time of the original work), an unsuspected insight emerged from the ratios of the successive frequencies in the curves of Figure 4. When the ratios of successive frequency values are computed and listed in sequence, i.e., F_1 / F_0 , F_2 / F_1 , etc., the values approximate a decreasing geometric series, the pattern being unbroken over both the rising and falling portions of the curve.

FIGURE 5 **Mathematical Form of the Curves of Figure 4**

GIVEN:

F_0	The frequency for 0-length chains
A	The ratio of F_1 to F_0
R	The ratio reduction factor at each step

$$F_1 = AF_0$$

$$F_2 = \frac{A}{R}F_1$$

- - -

$$F_k = \frac{A}{R^{k-1}}F_{k-1}$$

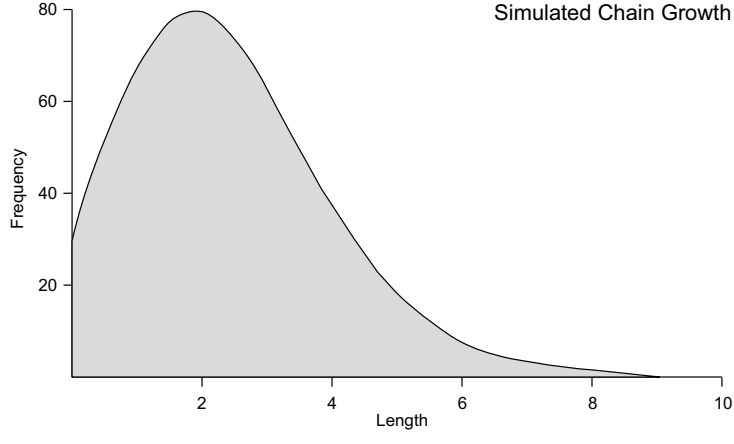
Furthermore, the ratio values decrease by about the same factor per step in all three languages. The mathematical formulation of this is shown in Figure 5. How are we to explain why the ratios decrease in this systematic manner,

leading to a precipitous, faster-than-exponential drop in the frequency values for the longer chain lengths? In keeping with the original motivation for the study, a statement in terms of the machinery of cognition would be in order, but none such has been forthcoming. Instead, a random process was found which can reproduce the chain-length frequencies quite well. The interpretation of this process and its parameters lies more in the domain of the evolution of languages, however, than in that of cognitive behavior.

The idea of a random process to generate the chain length data came up at first as the one assumed most unlikely. A little further thought, however, showed that it should be explored. Clearly, languages continue to evolve over time, so that in a period of perhaps a thousand years the vocabulary has undergone major changes. If the simple assumption is made that new words may be introduced with equal probability as 'offspring' of any existing word in the lexicon, and the growth process from an imaginary initial set of root terms is followed, there will be a period in which short chains are dominant and the population of longer chains is very small, just as in the actual data.

A programmed simulation of this uniform random growth, halted at a suitable point, produced the results shown in Figure 6.

FIGURE 6 **Random Growth of Chains**



Clearly the curve is similar to those of Figure 4, although the tail of this curve extends to greater chain lengths. Beyond the visual similarity, though, is the fact that the sequence of frequency ratios is close to a geometric progression but departs increasingly at longer chain lengths, giving rise to the extended tail.

It is generally held that the working vocabularies of actual languages do not simply expand indefinitely as new words are added. Older words are lost from common usage, presumably at a rate that, overall, matches the rate of addition. It is easy to extend the simulation to include such an obsolescence process, the only question being the form that it should take. The guiding principle is that, in equilibrium, every chain length must individually maintain a constant frequency. Thus we have the following relationship:

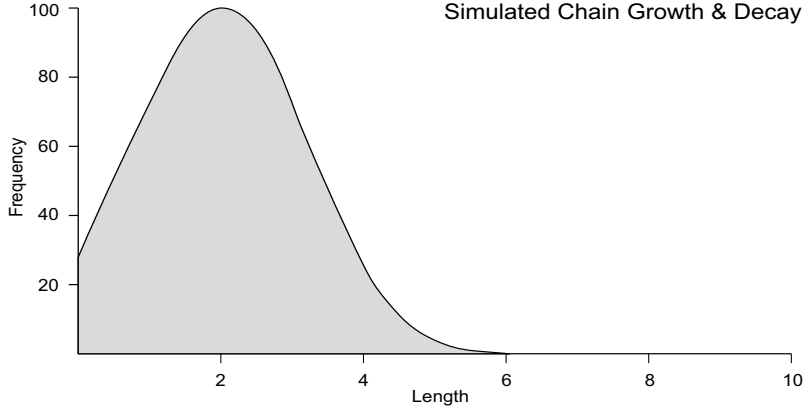
$$dF_n/dt = \alpha F_{(n-1)} - \Omega(n)F_n = 0 \qquad (1)$$

$$\alpha F_{(n-1)} = \Omega(n)F_n$$

$$\Omega(n) = \alpha(F_{(n-1)}/F_n) \qquad (2)$$

The constant, α, simply sets the time scale for the process, and the first term in Equation (1) represents the rate of addition to level n, which depends upon the *preceding* level, n-1, since new words are added as offspring of existing ones. Ω(n) represents the probability, relative to α, of the loss of a word at level n per unit time. Equation (2) shows that this relative probability of decay entirely governs the steady-state distribution. Accordingly, the simulation was extended to include an Ω(n) in the form of a rising geometric series. The resulting steady state is shown in Figure 7.

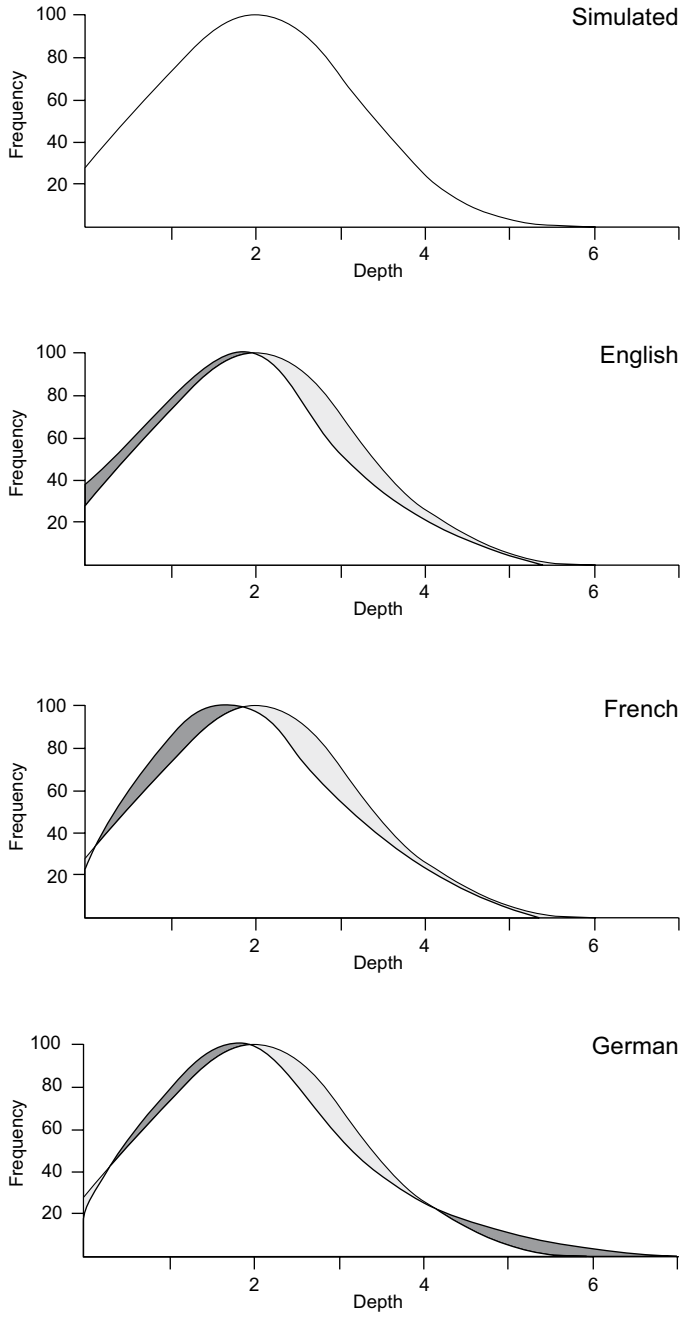
FIGURE 7 **Steady-State Chains**



Finally, in Figure 8 the data for the three languages are plotted together with the simulation data of Figure 7, showing a respectable degree of similarity among all four sets of data.

It appears likely, then, that the curious similarity among the chain-length statistics for the diverse languages is actually rooted in the universal process

FIGURE 8 **Simulated and Actual Data Superimposed**



whereby words enter and are eventually discarded from languages in general. The fact that the rate at which this proceeds appears to depend strongly upon the specificity of meaning, as represented by the definition-chain length, is of some interest, as is the quite regular form of this dependence. There is also an unexpected connection to Zipf in this, since he was quite interested in the process of coining words, which he saw, however, as being related to his main interest, i.e., word frequency.

It would be interesting to do an etymological follow-up, to see if, in fact, the age of words varies systematically with chain-length. The data here suggest that if the ages of words in a group sharing one chain-length are compared with those for the next shorter chain-length, the latter should be, on average, somewhat more than twice as old. This is because the probability of a word being lost to the group with chain length n is proportional to Ω(n), thus the average age of words in the group will vary inversely with Ω(n) (The geometric series of values for Ω(n) had a ratio of 2.25 in generating the data for Figure 7). Judging from the observed data, this age relation would probably be less exact for the longer chain lengths, five and above, particularly in German, where the increase in Ω(n) values seems to diminish considerably for the longest chains.

5.6 Brief Summary: A Logic-Problem Solver and the Cylinders System for Complex Structured Data

The preceding sections of this article have attempted to show how BCL's interest in human language evolved from the initial focus on pattern recognition and simulation of biological sensory systems, some of the early projects, and what conceptual framework was finally adopted in working with simulation of language behavior. In this final section, two projects of a more concrete nature will be summarized. Although the design of the Cylinder system came first, it will be more expedient to start the discussion with the problem-solver design.

During the period when BCL was active, computers became powerful enough to start performing tasks otherwise associated with intelligent beings, e.g., playing checkers or chess, solving algebra problems, or proving simple theorems in symbolic logic. It was natural, then, with our interest in language, to contemplate a program which could solve word problems of some complexity, such as the one which will be quoted below, which is similar to a number of such puzzles published by Lewis Carroll. This type of problem becomes quite complex when translated into the formalism of first order

predicate calculus, which was the only way that similar problems had been dealt with up to that time by artificial intelligence researchers, the solution being generated by adapting existing automatic theorem-proving techniques. From the BCL viewpoint, formal logic was not seen as the ultimate key to unraveling the mysteries of human language, so it was desired to find an approach that would work directly from a data structure intended to approximate the mental representation of the puzzle content. If that were possible, then the heuristic (i.e., 'rule of thumb') techniques already established in simpler, non-language-based, problem solving programs could be extended to this new material, and something much closer to actual human thought processes might be expected to emerge from the effort.

The chosen problem reads as follows:

A train is operated by three men: Smith, Robinson, and Jones. They are engineer, fireman, and brakeman, but not necessarily respectively. On the train are three businessmen of the same names, Mr. Smith, Mr. Robinson, and Mr. Jones. Consider the following facts about all concerned.

1) Mr. Robinson lives in Detroit.
2) The brakeman lives halfway between Chicago and Detroit.
3) Mr. Jones earns exactly $2000 annually.
4) Smith beat the fireman at billiards.
5) The brakeman's nearest neighbor, one of the passengers, earns three times as much as the brakeman, who earns $1000 a year.
6) The passenger whose name is the same as the brakeman's lives in Chicago.

Who is the engineer?

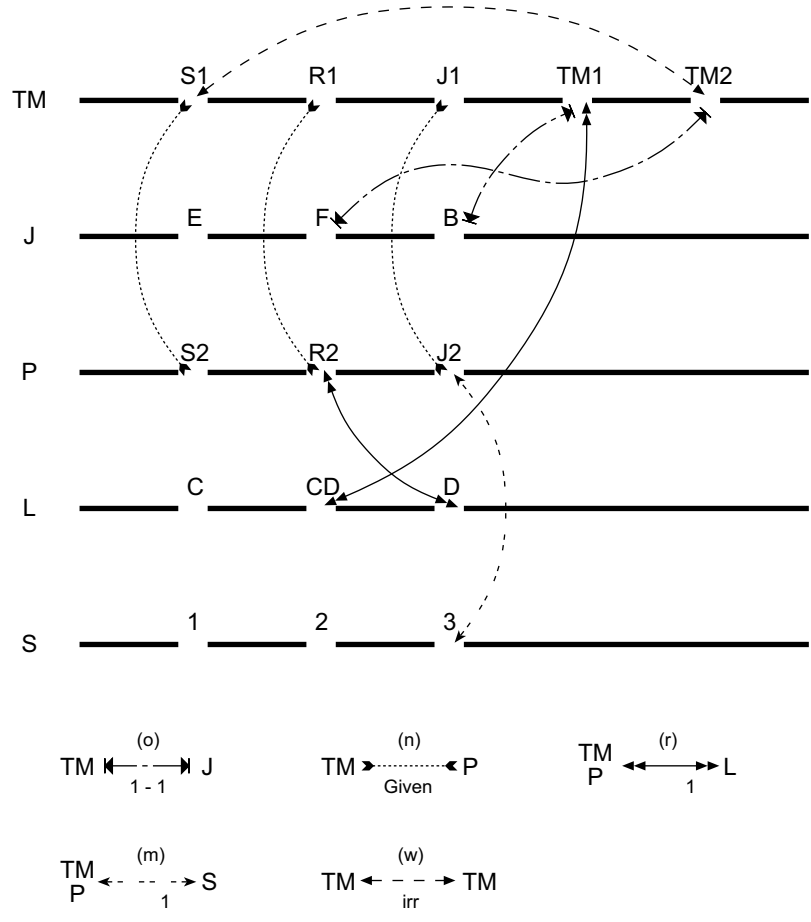
In setting up the representation of the problem content, it was seen that there were altogether fifteen people or objects, introduced in five sets of three: three trainmen, three jobs, three businessmen, three salaries, and three locations. There are five incomplete references to people, i.e., the brakeman, the fireman, the engineer (introduced by the final question in the problem text), the passenger who is the brakeman's nearest neighbor, and, lastly the passenger with the same name as the brakeman. The first three belong somewhere in the set of trainmen, the fourth and fifth in the set of passengers, but in each case it is not stated exactly which man is referred to.

In addition to the people and objects, the problem specifies five different kinds of relationship among them. These are: (a) having the same name, (b) having a given occupation, (c) living at a given location, (d) earning a given salary, and, (e) winning a game against someone.

The complete representation of the problem structure requires accounting for twenty people and objects grouped into five sets (the five partially-specified

people require separate representation in their respective sets), and a total of fourteen instances of the various relationships among these twenty people and objects. A diagram depicting the structure just after the representation for clue number (4) has been added appears in Figure 9.

FIGURE 9 **Partial Data Structure of "Smith, Robinson, and Jones"**



In the figure, the five sets are shown as horizontal lines with the elements listed in arbitrary order along them. The relationships are shown as links running between various objects and/or people in the sets; each relationship is shown in a different line and arrow style. At the bottom the logical properties of the five relationships are represented. These properties include the sets to which they pertain, whether or not a unique object or person is required

for either member of the relationship, and whether or not the relationship is explicitly defined in the problem context. The last relationship, [w], 'wins a game against', is specified as irreflexive, i.e., a person doesn't win against himself, a possibility that doesn't arise in the other four relationships, which relate elements from distinct sets.

The solution procedure involved an examination of the data structure to find, for instance, those unknown elements having large numbers of relational links, which could be expected to be promising starting points for a solution. The program then generated 'hypotheses', or tentative identifications of unknown elements with known ones, rejecting those which violated logical properties of the relationships. The hypotheses and their logical dependencies were recorded in a separate, much simpler data structure, having a hierarchical, or 'tree' form. Using straightforward rules for pruning the hypothesis tree to account for failed hypotheses, the unique solution was found in relatively few steps.

It was satisfying to see that such problems could, indeed, be handled directly within data structures representing the problem content, rather than necessarily having to be solved in more abstract formal logic. In addition, different kinds of questions could be answered with little change to the representational data structure. As an example, if the structure is augmented with the irreflexive relationship 'east of', the question "Does the passenger with the same name as the fireman live to the east of Mr. Smith?" can be easily handled. Or again, using the original structure, "What relation holds between Jones and the passenger with the same name as the engineer?"

It was less satisfying to realize that a great deal of real-world knowledge goes into the construction of the data structure from the problem statement. For instance, it is not in general impossible for a person to do two different jobs, although in the context of trains it would be very difficult to simultaneously serve in any two of the jobs mentioned in the problem. Similarly, people can have more than one home, but this particular problem has no solution without a one-to-one relationship between person and home.

The structured data shown schematically in Figure 9, which shows only about half of the relational connections that are involved in the entire Smith, Robinson, and Jones puzzle, offers just a glimpse of the complex, highly interconnected structures that are required to support problem solving by computer programs. As complexity grows, the interconnections quickly become difficult to represent clearly by means of a two dimensional picture. The Cylinder system was one practical answer to the problem of representing such multiply-connected structures within the very restrictive domain of computer memory.

The problem is that random access memory of a computer is essentially a one-dimensional structure; conceptually it is a very long row of 'pigeon holes' each capable of holding a small quantity of arbitrary data. Each such 'pigeon hole' is uniquely identified by the number representing its position in the row, i.e., its memory 'address'. This organization is adequate for storing such orderly structures as lists, tables, or, perhaps the sequence of characters in a text such as you are reading, but it seems to provide no way to handle data elements which have multiple, and arbitrary, connections to other elements.

The way to get around this comes out of two facts: first, that the computer's processor obtains its program instructions from the same memory system that holds non-program data, and second, that some program instructions must be able to specify arbitrary addresses in the memory system, since data to be processed by the program must be retrieved from the memory, and computed results returned to memory. Thus the storage capacity at each memory address must be large enough at least to hold an arbitrary memory address (plus a little more to indicate the program instruction to be performed). (It is true that modern CPU's are able to address memory in 'chunks' of several sizes, some of which are too small to satisfy the condition above, but there must be at least one size available which does meet the condition.)

Since there is room at each address in the memory to store another arbitrary address, this provides the means to make arbitrary interconnections among data elements, provided there are program instructions available to the machine's processor which allow non-program data to be interpreted as memory addresses. Even the earliest stored-program computers provided ways to use non-program data in this way, including Illiac I at Illinois, which was still in operation when BCL was organized. Not surprisingly, then, by the mid 1960's artificial intelligence researchers had designed a number of schemes using such stored addresses, or 'pointers' as they were called, to create data structures; in some cases entire software systems were built up around them, the most notable of these was probably the LISP language, designed at MIT.

The creation of Cylinders was motivated by the fact that the then-existing systems for structured data were basically two-dimensional in concept, while an inherently richer structure was, in our view, needed. Without going into fine detail, the minimal data-storage entity in the Cylinder system can be visualized as three-dimensional in structure, the internal linkage paths being representable, in general, as a wire-frame realization of a cylinder. The system offered rich possibilities for representing densely interconnected data structures, while simultaneously providing improved program efficiency in spite of the added complexity. The puzzle solver described at the beginning

of this section had, in fact, as one of its purposes the first fully functional test of the Cylinder concept.

That brings to a close this section, and with it these recollections of what we were doing in BCL. It has been a rewarding experience to walk again in that 'forest' and revisit some of the trees and shrubs; the forest died nearly forty years ago, but, in scattered locations, some of the trees are growing anew.