A QUITE GENERAL TEXT ANALYSIS METHOD

by

Dennis J. Clinkenbeard*

CU-CS-237-82 August, 1982

*Department of Applied Statistics and Computer Science, Utah State University, UMC 37, Logan, Utah 84322

ABSTRACT

Given a text, T, of some language, L, a method of analysis is described and demonstrated which will 1) search T for "meaningful" substrings of T and 2) using the "meaningful" substrings obtained, produce a finite automaton that accepts T. The procedure used to detect meaningful substrings relies entirely on a syntactic analysis of repeating strings within T. Thus, effectively nothing is assumed about T other than meaningful portions of T should repeat frequently if T is long enough. Subsequent to determining a collection of strings deemed meaningful, postions in T are equated in a manner that depends on their contexts with respect to the meaningful strings found. This ultimately provides a method allowing not only the creation of a finite automaton that accepts T, but also way of sorting the strings of T into grammatical categories. The combination of these two procedures yields a method sufficiently general to apply both to natural languages and to less understood languages such as the genetic codes of deoxyribonucleic acid (DNA) within biological organisms.

CHAPTER I

INTRODUCTION

Given a text T from some language, we describe a methodology together with an implementation to explore the internal structure of T. An investigation in this area is reasonable from a number of view points; to mention two:

a) no effective grammar has been given for any natural language and such grammars are likely to be necessary for any sophisticated language processing;

b) significant insights into the structure of genetic codes may be gleened if meaningful subunits of, for instance, genes can be discovered and categorized.

In particular, the strategy presented provides a means to more easily investigate the following observations, well known for natural languages, in a more general framework.

(i) Natural languages contain a hierarchy of "meaningful" units; e.g. paragraphs, sentences, phrases, words, nouns, verbs, morphemes.

(ii) Such "meaningful" units also tend to repeat more often than random strings of comparable length.

(iii) "Meaningful" units occur in different contexts; e.g. the English phrase 'running is a healthy exercise' is embedded in the sentence 'perhaps running is a healthy exercise but not exciting', and in the sentence 'gun running is a healthy exercise only if you are faster than bullets'.

(iv) Occurrences of "meaningful" units can be divided into categories according to the contexts in which they occur (Fries, [2]); e.g. if the sequence, 'the bobolok is hungry', were considered English then 'bobolok' would be a noun. Of course two occurrences of the same unit can be in different categories depending on their contexts; therefore any categorization is not solely on the units themselves but rather on their occurrences within texts.

The methodology we use is described in detail in the first section below. Essentially, we attempt to determine meaningful portions of a text T while assuming nothing more about T than that it is composed of a set of reoccurring strings whose arrangement is fairly restrictive. Then, using the occurrences in T of these presumably meaningful portions to define what is meant by the context of a position (in T), we equate positions that have similar contexts. This ultimately allows us to categorize substrings of T according to their contexts. Thus, assuming T is "long enough" so that a sufficient number of meaningful portions can be found, this method provides a quite

general way to explore the structure of texts.

To be more precise, we view equated positions as corresponding to the behavior of a finite automaton M cycling through it's states such that two positions are equated if M is in the same state at both positions. This, not surprisingly, leads to a regular grammar for T. Of course if the language containing T is not regular then M may be inordinately large and thus provide a rather distorted view of it's actual structure; however, even in such instances, some insight into the structure of the language is likely to be achieved.

It is also worthwhile to mention here that our method is similar to other text analysis techniques employed in linguistics. In particular, a similar strategy was successfully utilized in a semantic analysis of English texts by Zellig Harris [4].

In the second and third sections below, we give, respectively, an algorithm embodying our methodology and some features regarding it's implementation. It is important to note that, the algorithm was implemented such that the execution time bound is nearly proportional to the length of the input text. This was achieved by creating a data structure known as a position tree in which an efficient search for meaningful subunits of T occurs.

In the final section, we present some examples and results from investigations, conducted with the aid of the above mentioned implementation.

CHAPTER II

METHODOLOGY

In order to simplify our presentation and provide a more general framework, the following clarifications appear necessary:

(1) Any finite set of distinct symbols, E, will be called an alphabet.

(2) If T is a finite sequence of symbols from E and T is considered "meaningful" then T will be called a text. Although this is certainly not a rigorous definition, and indeed no rigorous definition can be given, the intent from a practical view point is to limit our investigation to sequences of symbols T such that, at least on some level, T can be conceptualized as a single functional unit; that is, there exists coordination or rules governing the arrangement between various portions of T such that T as a whole is viewed as something more than the sequence of symbols composing it.

(3) We define a language as an ordered pair (L,R) where L is a collection of texts and R is a collection of rules or grammar governing the arrangement of symbols within each T in L.

To illustrate the above notions consider the following two examples:

(a) If E={blank,'a','b','c',...,'z'} then the word 'boy' might be reasonably considered a text of the language English since the conceptualizations associated with 'boy' can be thought of as a functional unit and is normally understood as something different from the mere sequence of symbols: "boy". Moreover, a sequence of symbols such as: "the boy likes her" will also be considered as an English text even though the meaning is ambiguous.

(b) If E={'a','c','g','t'} such that these letters represent the four nucleotides composing genetic material, then a text may be defined as any string of these letters representing some biologically functional unit such as a gene(s) or chromosome(s). Thus, presumably since such texts obey certain structural rules, we will refer to the language associated with these texts as DNA.

This second example differs from (a) in that it is not known whether any of the four introductory observations about natural languages have analogues that are valid for DNA in general; however, at least for prokaryotic organisms, there are clearly defined contextual units delimiting genes; i.e. a Shine/Dalgarno region precedes each gene and one of the three stop codons terminates each gene.

In order to determine what the meaningful units of a given text or language are, or in deed whether the concept applies at all, we need a better understanding of what is meant by a collection of units

than English paradigms (such as the set of all English words, extended by syntax characters, as blanks, commas, and etc., or, the more elementary collection of all syllables also extended with syntax characters) provide. There are probably many ways to give a rigorous definition this concept, but since this is not our purpose, we instead give only the necessary technical descriptions to mitigate any concerns over ambiguity of the concept.

(4) Let S be a substring of a text T. If there exists a collection of strings C such that S can be constructed from some sequence of perhaps overlapping occurrences of elements of C then S is said to be covered by C. Furthermore, if there exists some way of arranging occurrences of elements of C (perhaps overlapping) such that their sequence exactly equals S then we say that S has as exact covering in C.

For example, if C= the set of all strings that have an optional blank at the beginning or the end and whose interior is an English word, plus, the set of all syntax characters, then the text, " this is exactly coverable." is exactly covered in C by sequence: " this ", " is ", " exactly ", "coverable", ".". However the sequence, "n example." is not exactly covered in C even though there is clearly a cover; e.g. "an ", "example", ".".

(5) In order for a collection C to be considered as a set of units for a language L the following properties appear sufficient:

(a) each member of C must repeat in different contexts within a substantial number of texts (e.g. we might specify that if u is a unit then there must exist an integer k, Ø<k<=1ØØ such that k percent of all texts of length <=N contains u in different contexts, N=1,2,3,...);

(b) a string s that repeats in each of a substantial number of texts is a member of C iff it does not always occur within another member of C;

(c) for each T in L there is an exact covering of T in C;

(d) units are not randomly arranged in any T; that is, there is a grammar for determining their order.

When attempting to either discover meaningful portions of a text T, or detect relationships between already known (combinations of) units, note that if one knows or assumes little about the semantics and internal structure of T then there are not many fruitful techniques to apply. About the only seemingly profitable notions remaining concern analyizing repeating substrings within T. A justification for this being that

(a) if there are indeed building blocks comprising T that are similar to the units described in (5);

(b) if meaningful portions of T correspond to strings exactly covered by the blocks; and,

(c) if T is sufficiently long then meaningful portions of T should not only reoccur but reoccur in discernible patterns.

Thus we make the following two assumptions:

(a1) repeats within T should give indications as to the structure and/or meaningful units of T;

(a2) longer repeats should give better indications.

Perhaps the most obvious approach along these lines is to attempt an examination of T by determining all repeating substrings and then analyize these substrings; however, such an approach appears overly naive at least when English text is considered. For example, consider the following text: "the derivative of the sum is the sum of the derivatives". There are clearly meaningful blocks that repeat in this text; e.g. "the derivative", " the sum ", " of the ". However, one must also contend with all repeats that are themselves inside another repeat; thus, such repeats as "rivative", "e sum" (and ultimately even single symbol strings as "a", "e", "i" and "s" if no minimum length is stipulated) are likely to interfere with any analysis.

To combat the deficiencies illustrated above we would like to do our analysis with only "maximal" repeats. Intuitively by a "maximal" repeat we mean a set U, containing at least two elements, with each element corresponding to a distinct occurrence in T of the same string, S, such that any attempt to lengthen S results in some member(s) of U not being able to be lengthened to produce an occurrence of the new, longer S. In order to make this notion precise the following slightly more general definition suffices:

(6) Given a positive integer n, an n-proper repeat in a text T is a substring S of T such that

(a) S occurs at least n times in T;

(b) For some n occurrences of S in T, S1,S2,S3,...,Sn:
  (i) if the symbol x1 precedes S1, x2 precedes S2, x3 precedes S3, ..., xn precedes Sn, then xi<>xj if i<>j, 1<=i<=n, 1<=j<=n;

  (ii) if the symbol y1 succeeds S1, y2 succeeds S2, y3 succeeds S3, ..., yn succeeds Sn, then yi<>yj if i<>j, 1<=i<=n, 1<=j<=n.

As an example, note that by taking n=2 in the definition and applying it to the problem text above, we retain the "maximal" repeats: "the derivative", " the sum ", " of the ", and, at the same time avoid many of the other less desirable substrings mentioned. In fact, the only other apparently disfunctional strings included in this

case are those having a length less than 4; these, of course, could be disqualified by requiring a minimal length of 4. Furthermore, note that with n=3 the only 3-proper repeats are single characters; i.e. "e", and "s". Thus as n increases the collection of "maximal" repeats decreases.

To illustrate proper repeats further consider the text:

he steadfastly denied going faster than anyone
else in the fast lane; even though admitting to
going as fast as the speed limit would be a fast
way to placate the officer.

The substring "fast" is a 2-proper repeat and also occurs as a substring of the 2-proper repeat " fast" which in turn occurs as a substring of the 2-proper repeat " fast ". Moreover, " fast " is also a 3-proper repeat.

In order to avoid confusion concerning the type of maximal repeat being discussed we forego using the word "maximal" and instead solely rely on the term n-proper repeat or proper repeat, specifying the value of n as necessary. Moreover, if no value of n is given we assume n is 2. We furthermore assume that the n-proper repeats being considered maintain some minimal length k which will also be specified as needed.

To give a more theoretical justification for our interest in n-proper repeats an operational definition of context is needed.

(7) Let S be a segment of the text T. By a context of S in T we mean an ordered pair of symbols (x,y) such that the segment xSy occurs in T.

Let L be a language with at least one collection of units satisfying (5c) above and let T be a text of L. If S is a non-nil repeat of T that does not have an exact covering, then S is not likely to be meaningful (although it may contain meaningful substrings). Therefore, with a higher probability than random, S either occurs in more limited contexts than strings of equal length that are exactly covered, or, is sufficiently short such that it appears frequently in T (perhaps in many different contexts) but is always contained in other units. Thus, by increasing the value of "n" for the type of proper repeat being considered and/or increasing the minimal n-proper repeat length, S can be filtered out.

Conversely, if T is long in comparison to the size of units, then their prevalence should not only result in many of them appearing in a variety of contexts but also other, longer, strings that are exactly covered. Hence the filtering process should be much less pronounced when the type of n-proper repeat is increased.

Therefore, as both "n" and the minimal length increase, the proportion of meaningful n-proper repeats should increase.

(8) To make precise what is meant by a position within a text T, we let g(i) denote the position within T between the i-1st and the ith symbol. For instance, if T="the boy likes grasshoppers" then g(1) corresponds to the position immediately before the "t", and g(3) corresponds to the position between the "h" and the "e" in "the". We shall call the g(i)'s gaps.

Assuming the above justifications for studying n-proper repeats are valid, we induce an equivalence relation R upon the gaps of T via the n-proper repeats of a fixed n. To accomplish this, the following property is used to generate R:

if (a) u is the longest non-nil n-proper repeat immediately preceding gap g(i) and v is the longest non-nil n-proper repeat immediately succeeding g(i), and,
(b) similarly, x is the longest non-nil n-proper repeat immediately preceding gap g(j) and y is the longest non-nil n-proper repeat immediately succeeding g(j),
then g(i)Rg(j) iff either u=x or v=y.

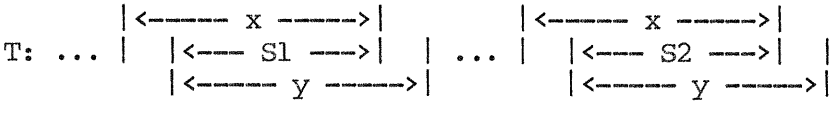
That is, the relation defined by this property is transformed into the equivalence relation R by taking the transitive closure.

For notational convenience, the equivalence class containing g(i) will be denoted by [g(i)]. If any [g(i)] contains more than one gap then we call such classes states, and g(i) will be called a state transition.

Using R, an equivalence relation ~ can now be placed on the substrings of T such that for any substring S of T if g(i) is the gap immediately before S and g(j) is the gap immediately after S then by mapping S into the ordered pair ([g(i)],[g(j)]) we obtain the desired equivalence relation on the substrings of T.

This new equivalence relation offers the following advantages:

i) if S1 and S2 are both occurrences of S in T, and x, y are both n-proper repeats of T which overlap giving S1 in one occurrence and S2 in another occurrence then S1~S2; that is, pictorially:



ii) if S1 and S2 are two occurrences of substrings in T, and x, y are n-proper repeats of T then if x and y provide the context for both S1 and S2 such that the substrings xS1y and xS2y occur in T then S1~S2.

Note that (ii) above is a standard linguistic technique for equating portions of text. Such portions are called replaceable.

Finally, it is important to note that such an analysis can proceed on various levels in a text depending on what is designated to constitute symbols in T. For instance, if T is composed of units similar to words as in English then the analysis can also be done using such "higher level" units as the symbols.

CHAPTER III

ALGORITHM DESCRIPTION

The first step toward doing the equivalencing of gaps above is to construct a position tree for the input text T. This data structure provides a convenient way to determine the position of an occurrence of any symbol, s, in T, essentially, by the sequence of symbols to the left of (and including) s. This is done by allowing paths in the tree to be labelled with substrings from T where the paths are of sufficient length such that each occurrence of a symbol in T is uniquely determined by a path from the root to a leaf in the position tree. A detailed exposition of position trees will not be given here; such an exposition can be found in either [1] or [3].

An algorithm using this data structure for discovering n-proper repeats within a text was chosen over more naive approaches since empirical observations indicate that position trees constructed on English texts grow only in proportion to the size of the text (see [1]). Therefore since we can find all n-proper repeats for a text by traversing it's position tree it appears that for our purposes position trees provide an efficient algorithm. Moreover, the naive approaches typically have a quadratic asymtotic time bound, which is restrictively slow on the lengthy input texts our method presupposes.

After the position tree is constructed, the longest n-proper repeat on each side of each gap of T is ascertained by traversing the position tree and employing the following algorithm:

```
(1) Proceed from the root down a path in the tree until
    the length of the path obtained is at least equal
    to the minimal n-proper repeat length the user has
    specified; let N denote the node reached; let S
    denote the string of symbols associated with the
    path from the root to N.
(2) If the node, N, encountered is not a leaf then
      begin(*looking for proper repeats*)
        if N has n sons then (*there are at least n*)
                             (*different occurences*)
                             (*in T of the string S.*)
          if there are n different symbols preceding
          some n occurrences of S
          (*note this is easily determined since the*)
          (*construction of the position tree requires*)
          (*fields within each node that give us this*)
          (*information immediately.*)
          then begin (*an n-proper repeat has been found*)
                 let r denote this n-proper repeat;
                 create a set, f, to contain each gap
                 that has r as the longest n-proper
                 repeat immediately to the right of it;
                 (*"f" will be called a "forward"*)
```

```
                (*proper repeat set.*)
                push f on a stack of "f-sets", FSS;
                create a set, b, to contain each gap
                that has r as the longest n-proper
                repeat immediately to the left of it;
                (*"b" will be called a "backward"*)
                (*proper repeat set.*)
                push b on a stack of "b-sets", BSS;
              end;
      traverse the subtree below N;
      if elements were pushed on stacks FSS
        and BSS above
      then pop FSS and BSS;
   end(*looking for proper repeats*)
Else(*N is a leaf and therefore the path to N*)
    (*uniquely identifies some gap by some string*)
    (*of symbols that follow it.*)
  begin(*a leaf found*)
    let g(i) denote the gap uniquely determined by N;
    if FSS is not empty then(*the top element of FSS*)
                    (*corresponds to the longest*)
                    (*n-proper repeat, r, to the*)
                    (*right of g(i)*)
       begin(*proper repeat found*)
           add g(i) to the f-set, f, at the top of FSS;
           (*For each proper repeat p let j(p)*)
           (*denote the length of p. If r(f) is*)
           (*the proper repeat associated with f then*)
           (*certainly the gap g(i+j(r(f))) has r(f)*)
           (*immediately to the left of it.*)
           (*However, note that for each proper*)
           (*repeat, s, such that s corresponds to*)
           (*a b-set currently in BSS, the gap*)
           (*g(i+j(s)) also has an n-proper repeat*)
           (*immediately to the left of it. Therefore*)
           (*we must check to determine whether s*)
           (*is longer than any other proper repeat*)
           (*that has previously been found to the*)
           (*right of g(i+j(s)).*)
           for each b in BSS do
             begin(*checking b-sets*)
                let m be the length of the n-proper
                repeat associated with b;
                (*i.e. m:=j(r(b)) *)
                if g(i+m) is not in any other b-set
                   then add g(i+m) to b
                else if m is greater than the length
                      of the b-set currently containing
                      g(i+m) then
                      begin
                        delete g(i+m) from current b-set;
                        add g(i+m) to b;
                      end;
```

```
                end;(*checking b-sets*)
            end;(*proper repeat found*)
        end;(*a leaf*)
(3) Go to (1) until each node has been encountered in
    (1) or (2) above.
```

In estimating a time bound on the execution of (1) through (3), recall that the expected size of the position tree is proportional to the size of the input text (in fact, for texts in either DNA or English observations indicate 2 to 2.5 nodes per input symbol). Thus, the description above is linear with the exception of the loop to check b-sets. This loop has the potential of increasing execution time beyond linearity. But if the assumption is made that for at least all large input texts T, the probability is negligible of repeats occurring whose lengths are greater than M*log(K) where K=length of T and M is some fixed constant independent of T, then the length of most paths in the position tree is bounded by this expression. Therefore since BSS can have no more elements than the length of the path from the root to the node currently encountered, the b-sets loop can boost our asymtotic time bound by no more than a log(K) factor on such large texts. Hence, the order of our algorithm thus far is no greater than K*log(K).

Moreover, it should be noted that if a second position tree were constructed to uniquely determine each position in T by sequences of symbols to the right (instead of to the left as described above) then, by traversing this new tree, "backwards" n-proper repeat sets could have been detected in a manner identical to the way "forward" n-proper repeat sets are detected above. Hence, the b-set loop could have been eliminated above, yielding a linear algorithm. This alternative algorithm was not seriously considered due to (1) difficulties implementing position trees, some of which will be mentioned in the succeeding section; and (2) observations indicating that whenever a leaf is encountered in the position tree, the b-set loop is seldom iterated more than twice.

When the above portion of the algorithm is completed each gap is potentially in one f-set and one b-set. The gaps must now be transitized to create the equivalence relation R described in the previous section. The following is a brief description of the algorithm used.

```
1) Find the next gap, g(k), from the beginning of the
   text that has not been equivalenced into some [g(i)]
   but such that g(k) is contained in either a b-set or
   an f-set; If g(k) is found
      then let Z be the (f or b)-set containing g(k)
   Else let Z be nil;
2) If Z is not nil then
    begin
       create a new equivalence class [g(k)] in which
       to insert gaps;
       repeat
          add the gaps in Z to [g(k)];
```

```
        if Z is an f-set  then
           for each g(m) in Z do
            begin
             if g(m) is in a b-set, b, and b is not
                a subset of [g(k)] then
                  add b to the set U which contains
                  all b-sets and f-sets whose gaps
                  have not been examined as those of
                  Z are currently being examined
            end
         else (*Z is a b-set so*)
              for each g(m) in Z do
                 if g(m) is in an f-set, f, and f is
                    not a subset of
                    [g(k)] then add f to the set U;
          Z:=same (b or f)-set in U; delete Z from U;
      until U is empty;
      go to (1);
    end.
```

It should be noted that given appropriate data structures this portion of the algorithm can (and was) implemented such that the asymtotic time bound is linear (see appendix 4 for program listing).

Finally once the gaps in T have been equivalenced, the equivalencing of substrings in T is straight forward; however, the question arises as to what substrings are "interesting" enough to continue with any further processing, or to output. Clearly only a small portion of the exponential number of such substrings can continue to be considered. The implementation contained in appendix 4 provides options for: 1) reprinting the text with integers interspersed specifying the equivalence class of each gap in T that we called states in the methodology section above; 2) printing each n-proper repeat r followed by ordered pairs of integers designating the ~ equivalence classes in which the occurrences of r are contained; and 3) printing the members of the ~ equivalence classes which are also n-proper repeats. An illustration of these options, on a short English text can be found in Appendix 1.

CHAPTER IV

DESCRIPTION OF IMPLEMENTATION

Here we give a brief description regarding some of the features and difficulties entailed in the implementation of the algorithm above. The program is written in PASCAL; this language was chosen not only due to the fact that it's flexibility in creating data structures and relative efficiency seem to provide an adequate environment for subsequent program additions and modifications, but also, due to the familiarity it has obtained among a large group of potentially interested users.

A primary design concern was that the implementation should be capable of accommodating an alphabet of at least 50 symbols. This stipulation dictated a design where, essentially, there is no inherent restriction on the number of symbols the alphabet can contain. Thus, words or other equally large sets may be considered as an alphabet by merely changing the definition of symbol. However for large alphabets an internal encoding to some data type such as integer would be advisable to decrease both the program working set and the total storage necessary during execution.

Initial designs for implementation seemed to indicate that the crucial factor would be the amount of over all memory used during execution; therefore the first implementations reflected this consideration by making most data structures dynamic. However, when executing these versions upon even moderate input texts (5000 to 7000 symbols) on a VAX 11/780 with 2 megabytes of primary memory and UNIX operating system, the actual execution time increased in a quadratic fashion due to the excessive number of page faults resulting from the relatively large working sets dynamic allocation necessitated. Subsequent versions have mitigated this difficulty by transforming many of the previous dynamic data structures into static counterparts, albeit, increasing the over all execution storage requirements.

A significant amount of thrashing also seemed to be a direct result of using a position tree as one of the central data structures. The problem being not only that large tree structures are often not amenable to small working sets when dynamically created nodes are scattered throughout memory, but also that, in our particular use of position trees, even the fields within the nodes themselves are not easily confined in small portions of memory when the number of symbols that can appear in the text make it prohibitive to statically declare a single maximum size for all nodes. This latter handicap was lessened by implementing a feature capable of writing the position tree to files periodically during construction and then reconstructing the tree from these files so that at least nodes of the tree become localized. Furthermore, as an added benefit, this feature allows position trees for large input texts to be stored in files for reuse in later executions.

The following features have also been added to facilitate investigations:

(1) A special character can be given which if encountered within the input text, the program is prohibited from identifying occurrences of n-proper repeats that contain this special character. This allows multiple texts to be analyzed together in a single execution without perhaps spurious n-proper repeats being obtained from the tail of one input section and the beginning of another.

(2) During a single execution, analysis of the input text can be iterated with both the minimal n-proper repeat length and the "n" on the type of n-proper repeat varying according to user specified ranges.

CHAPTER V

RESULTS AND OBSERVATIONS

Our results demonstrate that this procedure provides a means of determining some of the meaningful strings in a text. For example, consider the following text:

> the japanese, well known for their ingenuity and adaptability, continue to receive international attention due to their growing economy and technological advances. since the end of the occupation in 1954, japanese gross national product has grown on the average of 9.6% annually.
> they are now second only to the united states in gross national product producing more steel, ships, and automobiles than any other country in the world. japan is also considered to have the most advanced train technology in use in the world.
> development has been fast and powerful for the japanese and many of the problems in the economy may be attributed to change occurring too fast for either planning or mitigation. some of the obvious difficulties for the japanese recently have been not only the lack of natural resources and raw materials, but also extensive pollution and the ill health of people due to the high population density in industrial areas. concern for these and other problems can be seen historically, but the real efforts to deal with such problems have occurred mostly since 197Ø. the solutions originated at the demands of the people most effected. policy and legislative debates brought publicity to the questions and encouraged even more public involvement.
> a glance historically at societal trends and actions offers insight into the progression of the problems, current attitudes, and solutions.
> the meiji restoration of 1868 introduced industrialism as a means to world power for the japanese. the constitutional revolution, as it is called, uprooted the feudal system in effect until that time and established rights and privileges for all japanese. the new constitution instituted unprecedented reforms that set the country on its new economic path toward industrial development.
> technologies and industries most heartily supported were the train system, communications systems, fabric manufacture and mining operations. the economy flourished rapidly after these changes.
> considerations of what the limits of rapid development might be and the possible negative influences

of unchecked growth were not thoroughly examined by the government or the people. some early incidents involving the emission of noxious gasses into the air and poisonous substances into water tables in the 189Ø's were not given proper attention. only in extreme cases where health hazards and immense public pressure forced the polluters to improve conditions did companies make changes. in early times these changes usually involved compensation for the injured party rather than actual elimination or mitigation of the problem. any changes were prompted by government policy though it was weak and incomplete for a long time. this was due to a strong alliance between business and government and the historically maintained proposition that whatever is best for industry and development was best for the people and country. thus, it was common for policy implementation that involved restriction of industry to be thwarted. the pollution problems were much compounded by the tremendous migration of the people from rural and agricultural areas to the city. everyone was attracted to the city center where jobs and prosperity were abundant. not until the mid-196Ø's did the government make unignorable laws that forced industry to plan for and prevent environmental damage. from the 195Ø's to the 197Ø's pollution problems arose and received attention, but the first law defying the trend of government's uninvolvement in commercial pollution came in 1967 with the basic law for environmental pollution control.

the basic law clearly defined environmental standards, necessary pollution prevention programs, and assigned the offender complete responsibility for controlling pollution.

thus, a growing awareness that anti-pollution regulations must be harmonized with national economic growth had taken root. in 1969 the government enacted a relief for victims of environmental pollution law that began movement towards strict control of pollution. this law, though unprecedented on paper in other countries, dealt with pollution after the fact and was not immediately matched with energetic enforcement. under the pollution relief law, the costs of pollution relief are shared by the offender, national government, and local government. criticisms of this system of relief are that the entire system is under government control with no public input in the selection of physicians or in the determination of victims.

the court process for settling cases is long and expensive and only an estimated few of the most serious types of injuries actually even attempt to

receive compensation. another major step in legislation took place in 1970 when 14 amendments to the basic law were passed applying effluent standards to all bodies of water and ambient air quality standards to the air. the new laws also gave more strength to the enforcement of the standards and provided substantial penalties for non-compliance or violation of the established standards. it then became obvious to japanese officials that a ruling body must be responsible for the enforcement of laws and the evaluation of specific problems. the japanese environmental agency became that ruling body in 1971. it consolidated all the various parts of government dealing with pollution related problems.

after its establishment, several key cases were brought before the courts and resulted in substantial financial penalties for the offenders. despite this display of authority, many feel that the japanese environmental agency suffers from such a lack of legal and political power that it cannot really enforce its mandates. for example, japan environmental agency has no authority over public utilities or services. moreover many environmental and pollution concerns are currently neglected by other agencies, ministries and industries whose environmental impact is critical. these claim that affirmative action in this area is now the responsibility of the japanese environmental agency.

despite these problems, the japanese have continued their efforts to improve environmental quality by amending these basic laws. some of the most critical and influential laws have been the air pollution and water pollution control laws of 1962 and 1971, respectively. the air pollution control law was the first law in this field and established procedures for control of emissions of soot and smoke from factories, but actually had little immediate effect on air quality. the water pollution control law, though coming much later, was more effective. it prompted the designation of standards for 81% of the public water ways of japan. the sunshine project of 1974 is a research and development program engaged in the study of potential effective uses of solar, wind, hydrogen, geothermal and other less polluting sources of energy. comprehensive measures have been adopted in land use planning to counter the view that the value of land is based on its development potential. green belt buffer zones for urban areas are being provided to relieve congestion. in addition, the nature conservation law of 1972 protects wildlife and natural habitats in some areas of japan. overall, land use classi-

fications are assigning value to resources balanced between development and preservation. thus, more emphasis is being placed on multiple uses of land.

When program parameters were set such that 3-proper repeats with minimal proper repeat length 8 were asked for, the following strings were output:

" attention", " control", " control ", " development ",
" environmental ", " government ", " government",
" the government ", " japanese", " japanese ",
" the japanese", " the japanese ", " national ",
" policy ", " pollution ", " pollution control",
" problems ", " problems", " public ", " relief ",
" standards", " standards ", " the offender",
" the basic law ", " and the ", " for the", " for the ",
" in the ", " of the ", " that the ", " to the ",
"ed by the ", "n for the", "ation of ", "e public ",
" of the p", "s to the ".

Note that of these 37 strings, all but the last 6 are clearly meaningful. Moreover of the exceptions, "ed by the" and "ation of " might be considered reasonable representations of grammatical categories. For instance, in all cases in the text "ed by the" provides sufficient information to identify the following pattern: (verb root)"ed by the "{adjective}(noun) where enclosures within curly brackets denote optional terms. Indeed, the following phrases can be found in the text: "shared by the offender", "examined by the government", "compounded by the tremendous migration". In the case of "ation of ", we can identify the following pattern: (verb root)"ation of"(noun phrase); i.e. the text contains: "designation of standards", "determination of victims", "evaluation of specific problems", "migration of the people", "mitigation of the problems", "restoration of 1868", "violation of specific problems". Thus, even for this fairly short passage only four of the 34 strings do not correspond to a meaningful unit. Furthermore, it bears mentioning that this collection certainly contains many of the key words of the text.

Also, note that if we relax either of the parameters a high percentage of less desirable strings are also included in the output. For instance, when 3-proper repeats of minimal length 7 was asked for these additional strings were also provided:

"al and ", " and in", " and po", " and pr", " areas ",
"ations ", "ation. ", "effect", "ensive ", "es for ",
"lation ", " other ", " system", "s that ", "s were ",
" the co", " the de", " their ", " these ", " water ".

When 2-proper repeats of minimal length 8 was given as parameters well over 100 additional strings were added to the original list above, and approximately half of these additional strings would not be considered meaningful.

As a final comment on the above example, the categorizations of substrings in the text by the state transitions are somewhat different

from standard grammatical classifications. For instance, there is a state transition in the text between " of the p" and "roblems" and the same state transition occurs between " of the p" and "eople", and, as another example, there is a state transition between " for the" and "se". However, this second example does provide some insight into how a finite automaton might process the text. Since " for the" also occurs as the initial portion of the string " for the ", the state transition after the "e" can be viewed as an intermediate state from which two different grammatical configurations are possible.

Also it is not unusual to see little, if any, equivalencing on occurrences of the proper repeats when a sufficiently long text and reasonable parameters are given such that apparently meaningful proper repeats are found. Thus, no occurrence in the text above of a 3-proper repeat from the initial collection was equivalenced to an occurrence of a different string on this list. This turns out to be a mixed blessing. On the one hand, the largest set of meaningful proper repeats that can be obtained by adjusting the parameters, thus far, has always been relatively small in comparison to the text. Since the occurrences of the proper repeats in the text ultimately determine the categories of all other strings in the text, to get as wide a range of categories as possible, one or two proper repeats in a category does not seem unreasonable. But on the other hand, from the initial list above one might expect any reasonable categorization to put " for the ", " of the ", and " to the " in the same category.

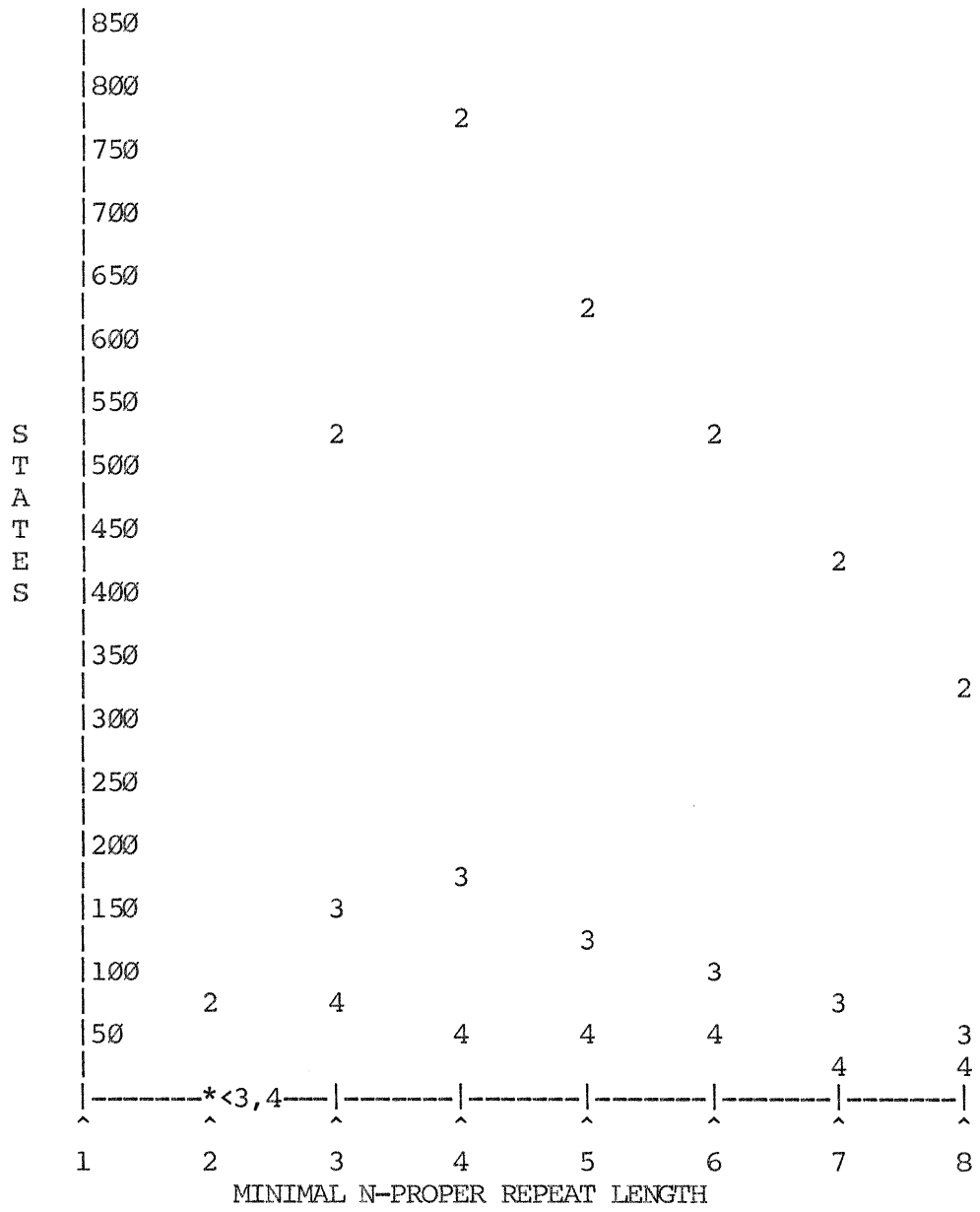
It is important to keep in mind that the primary advantages of our procedure are it's simplicity and generality. The program has no provisions for determining what strings constitute words or phrases; all symbols are essentially treated the same. Thus when the output is the input text interspersed with the state transition (if any) of each gap as in appendix 1, it is interesting that most states are between words (on the order of 85%) if 3-proper repeats of minimal length 8 are specified as in the above example. Moreover, those states that occur within words are likely to be on the boundaries between morphems.

As an example of this phenomenon, in Appendix 2 we give a longer text (than above) interspered with designations for the state transitions, plus, a list of the n-proper repeats found. Note that although the text is much longer, some of the state transitions still seem to be intermediate or spurious states; for instance, there is a transition state between "a" and "nd" of the word "and", between "p" and "ollution" of the word "pollution", and between "na" and "tion" of the word "nation".

Even though the equivalencing of strings in texts does not currently seem to provide sufficient information to analyize the text in any detail, the number of states obtained, as we vary the minimal length and the type of n-proper repeat, supplies some useful information in determining if units exist and a measure of their size if they do. The subsequent paragraphs furnish the justification for this statement.

Let i and j be two integers such that i<=j. If an n-proper repeat, r, is found in T with the minimal n-proper repeat length j then certainly r will also be found when i is the minimal length. Furthermore, it is not difficult to show that for each gap g, the state containing g (if any) produced when j is the minimal length is contained in the state (if any) produced when i is the minimal length. Assuming there are sufficient n-proper repeats of minimal length j such that for most states, s, obtained with minimal length i, there are at least n gaps g in s with an n-proper repeat of length j on one side or the other of g, then a large equivalence class of gaps produced with minimal length i will be divided into a number of smaller equivalence classes when j is given as the minimal n-proper repeat. Thus, under these circumstances, the number of states associated with i will be less than the number of states associated with j. Of course, as j increases, eventually the number of n-proper repeats decreases sufficiently such that some states that exist when the minimal length is i, will disappear completely at minimal length j; hence, the number of states ultimately becomes zero. Therefore for each T there is an interval of minimal length values (likely to be a single value) that yield the maximum number of states.

As an illustration, the data on the following graph was produced when the childrens book, Red Man, White Man, African Chief, (see Appendix 2 for text) was provided as text.



Graph of the states as a function of the minimal n-proper repeat length for the text, Red Man, White Man, African Chief.

2 denotes states for 2-proper repeats.
3 denotes states for 3-proper repeats.
4 denotes states for 4-proper repeats.

Notice also that the maximum values on this graph occur in the range one might expect would be likely lengths for syllables (when extended by syntax symbols). In fact, we can give an argument that lends some justification to this comment.

Let m be the minimal n-proper repeat length and k be the average length of the occurrences of units in a text, T, which can be exactly covered by these units. If m is sufficiently small, such that almost all units are longer than m, then upon processing T it is likely that

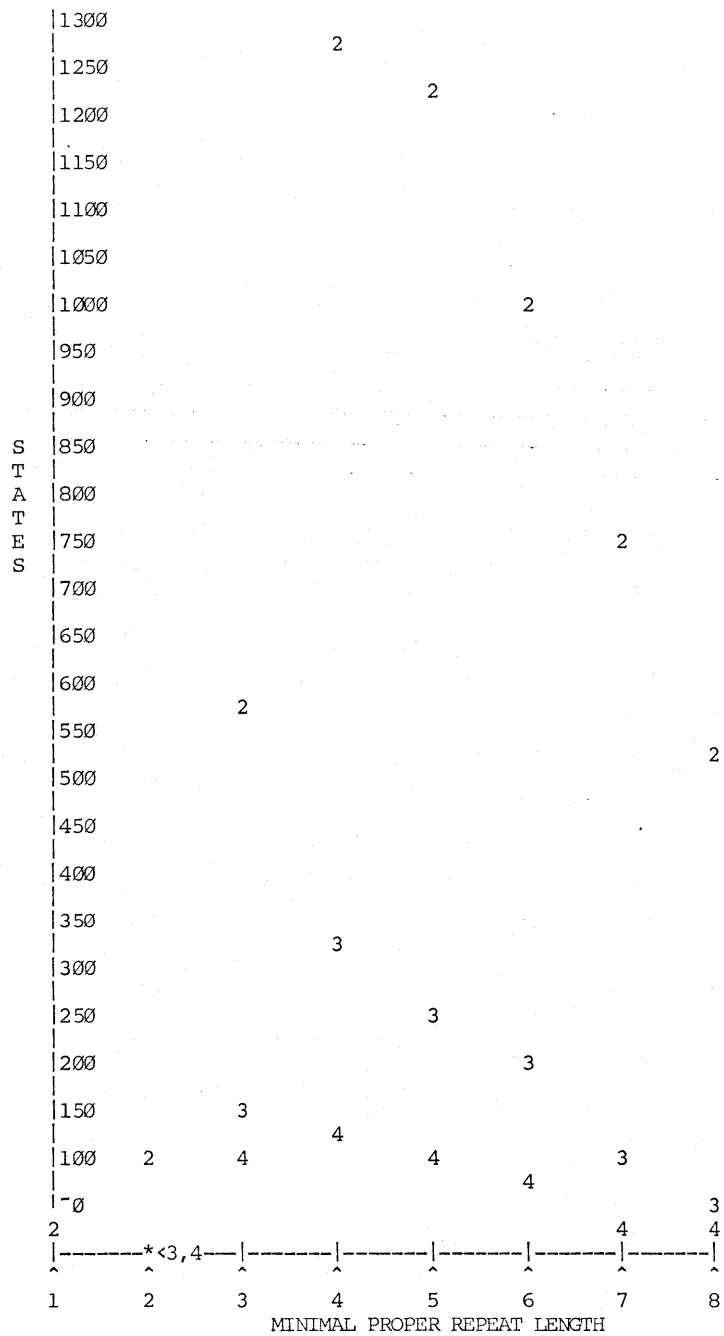
(a) n-proper repeats, r, of length greater than or equal to m and less than or equal to k, will be found that can not be exactly covered;

(b) most n-proper repeats that can be exactly covered will also be found. Assuming most such non-exactly coverable r's, due to their shortness, appear frequently in T, it is also likely that most of them have an occurrence adjacent to at least one longer n-proper repeat that is exactly coverable. This implies that any state containing n-proper repeats such as r, also likely contains longer n-proper repeats that are exactly coverable. Therefore most states obtained with minimal length m will exist when the minimal length is increased by one. In fact the total number of states should increase as was indicated in one of the paragraphs previous to the above graph.

This analysis implies that if a text is exactly covered by units then the smallest minimal n-proper repeat length, M, that yields a maximum number of states is likely to be a guaranteed length where a significant number of units are shorter than M. Thus M can be viewed as a measure of the length of units.

Furthermore, note that since our reasoning is based upon the fact that the r's are short and frequent, the greater the ratio these are in comparison to longer, or less frequent non-coverable n-proper repeats, the better our reasoning should apply. Now as the value of n increases, this ratio ought to also increase. Therefore, M should become a better measure.

So that this line of reasoning seems more plausible, we also present the following graph which describes the state --- minimal n-proper repeat fluctuations for the sample text presented at the beginning of this section.



Graph of the states as a function of the minimal n-proper repeat length for the sample text at the beginning of this section.

2 denotes states for 2-proper repeats;
3 denotes states for 3-proper repeats;
4 denotes states for 4-proper repeats.

To shed some light on the question of whether a given text contains units or not, we will consider the extreme cases; that is, English where units exactly cover each text and randomized English texts where presumably no units exist. In comparing long English texts with their randomized versions, certainly the English texts should seemingly have a higher frequency of long n-proper repeats. But how much higher a frequency and at what lengths might lead us to suspect one has an abundance of units and the other none? This question could likely be answered in some manner by statistical methods; however we illustrate a relatively simple way to distinguish between the two.

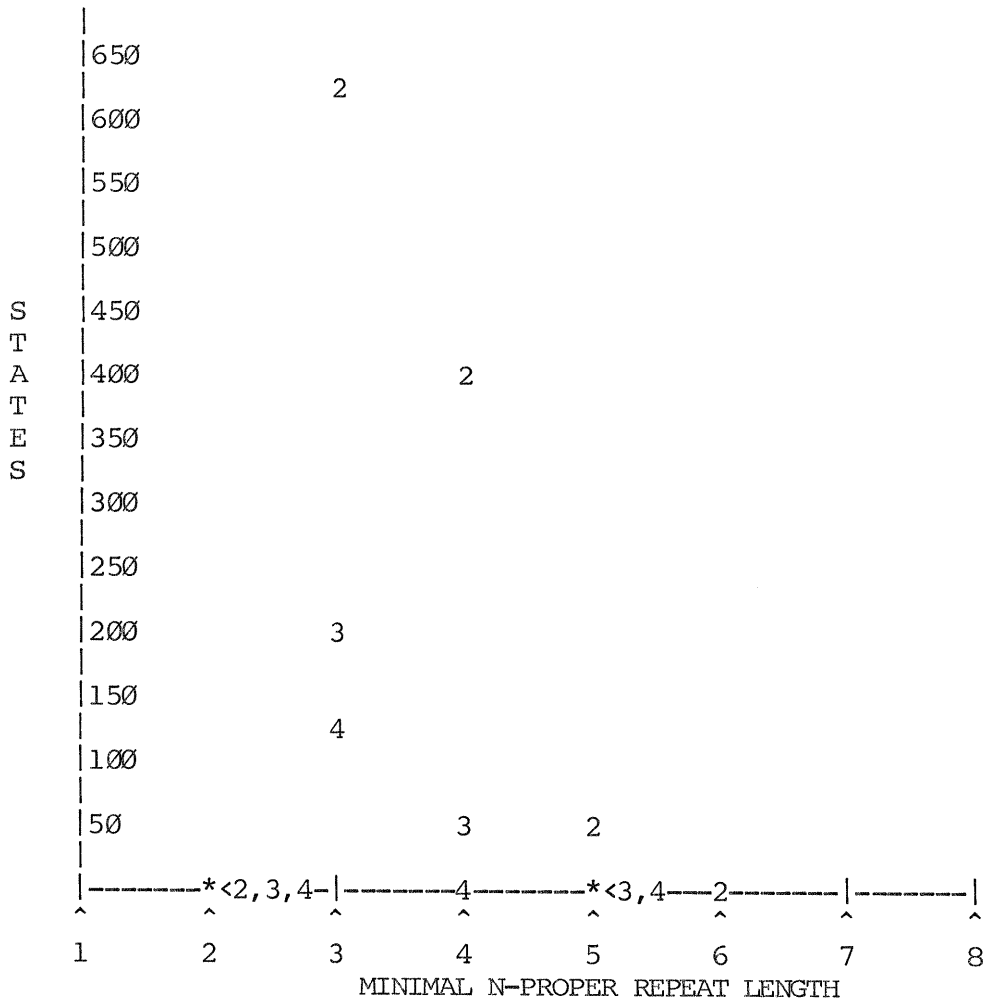
Let us assume that at least for English texts the following two propositions hold:

(A) Our comment above about measuring the length of units is true;

(B) For the minimal n-proper repeat lengths (of fixed n) that are at least twice the value of M (i.e. at least twice the value of the minimal n-proper repeat length yielding the maximum number of states), each state corresponds to a set of gaps whose equivalencing is due to only one n-proper repeat. (Note that although this assumption may appear fairly restrictive, no input thus far processed by the program has yet violated it.)

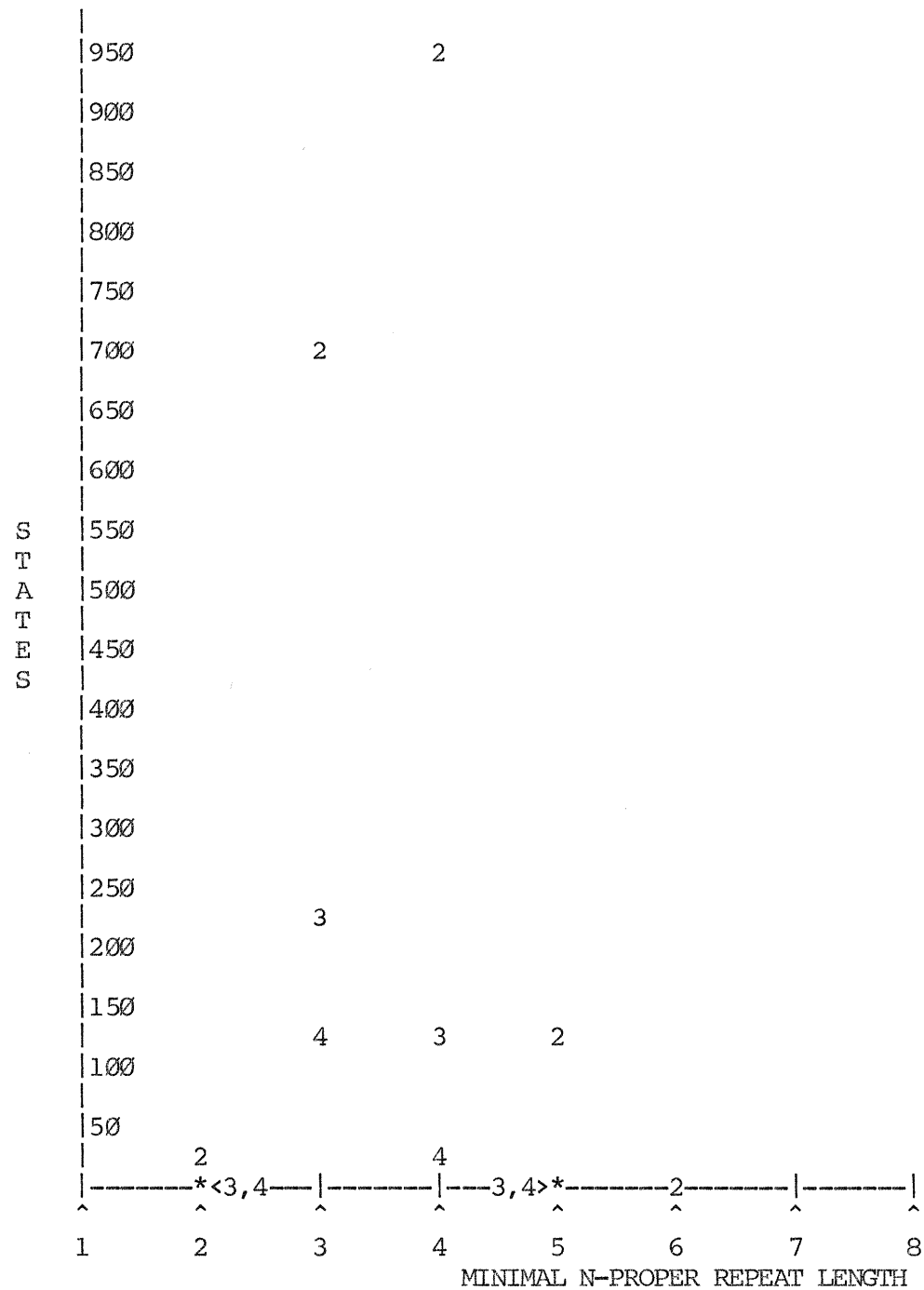
Let T be an English text. By (A), a significant number of units have a length less than or equal to M. If most of these small units have (for example) a length of M-2 or greater then n-proper repeats of length approximately twice M are likely to be found since this would be close to length of two adjacent units. Therefore by (B) above, a graph of the states should yield a non-trivial number of states when the minimal n-proper repeat is 2*M.

This is certainly illustrated by the plots above. However, the graphs for randomized English do not display this phenomenon. As typical examples consider the following two graphs. The first graph is for the randomized text of the children's story, Red Man, White Man, African Chief. The second is for a randomized version of the sample text at the beginning of this section.



Graph of states as a function of the minimal n-proper repeat length for the randomized text of Red Man, White Man, African Chief.

2 denotes states for 2-proper repeats;
3 denotes states for 3-proper repeats;
4 denotes states for 4-proper repeats.



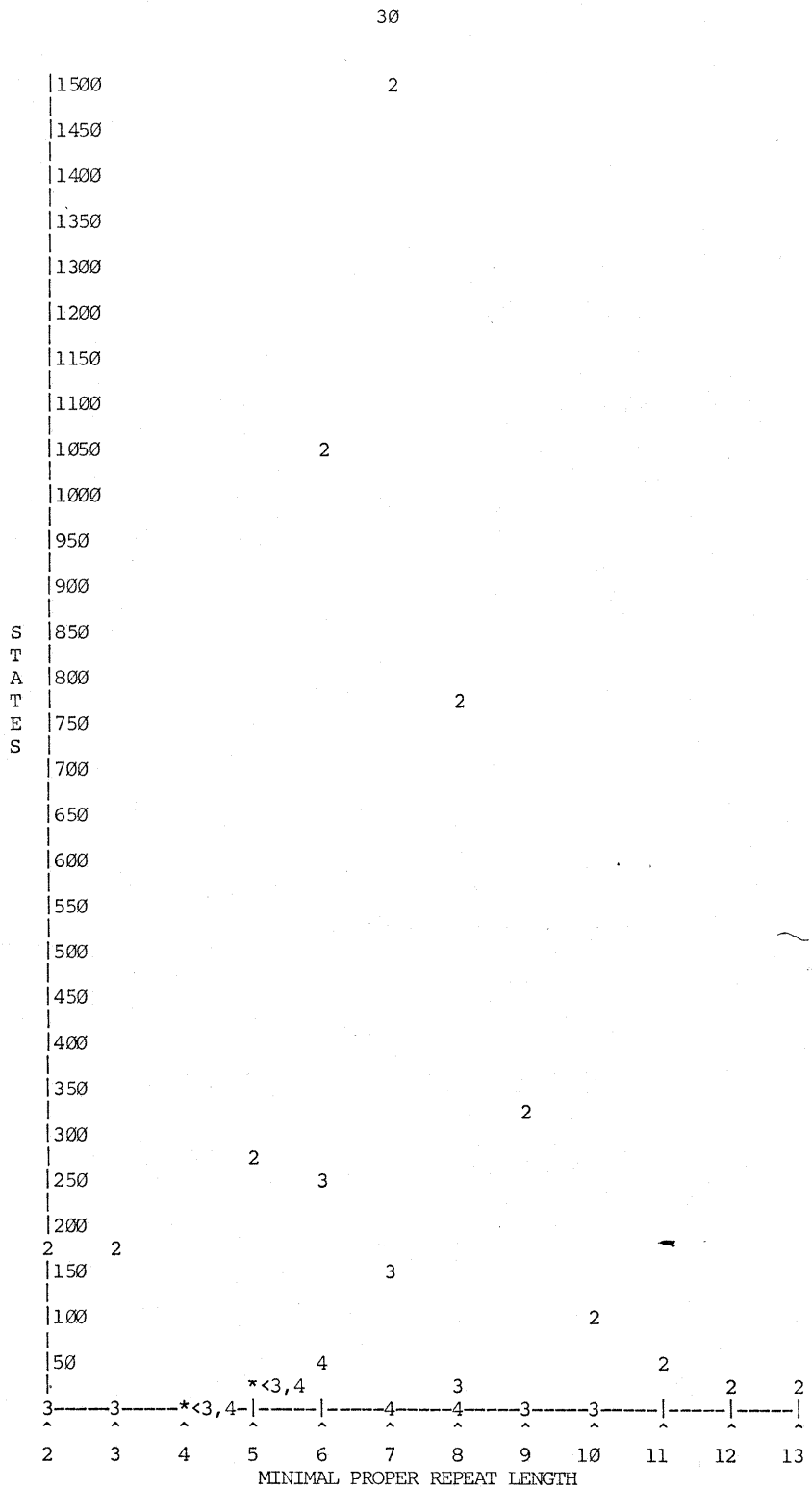
Graph of the states as a function of the minimal n-proper repeat length for a randomized version of the sample text at the beginning of this section.

2 denotes states for 2-proper repeats;
3 denotes states for 3-proper repeats;
4 denotes states for 4-proper repeats.

We have thus far mentioned nothing specifically concerning the language DNA; therefore some comments are probably warranted.

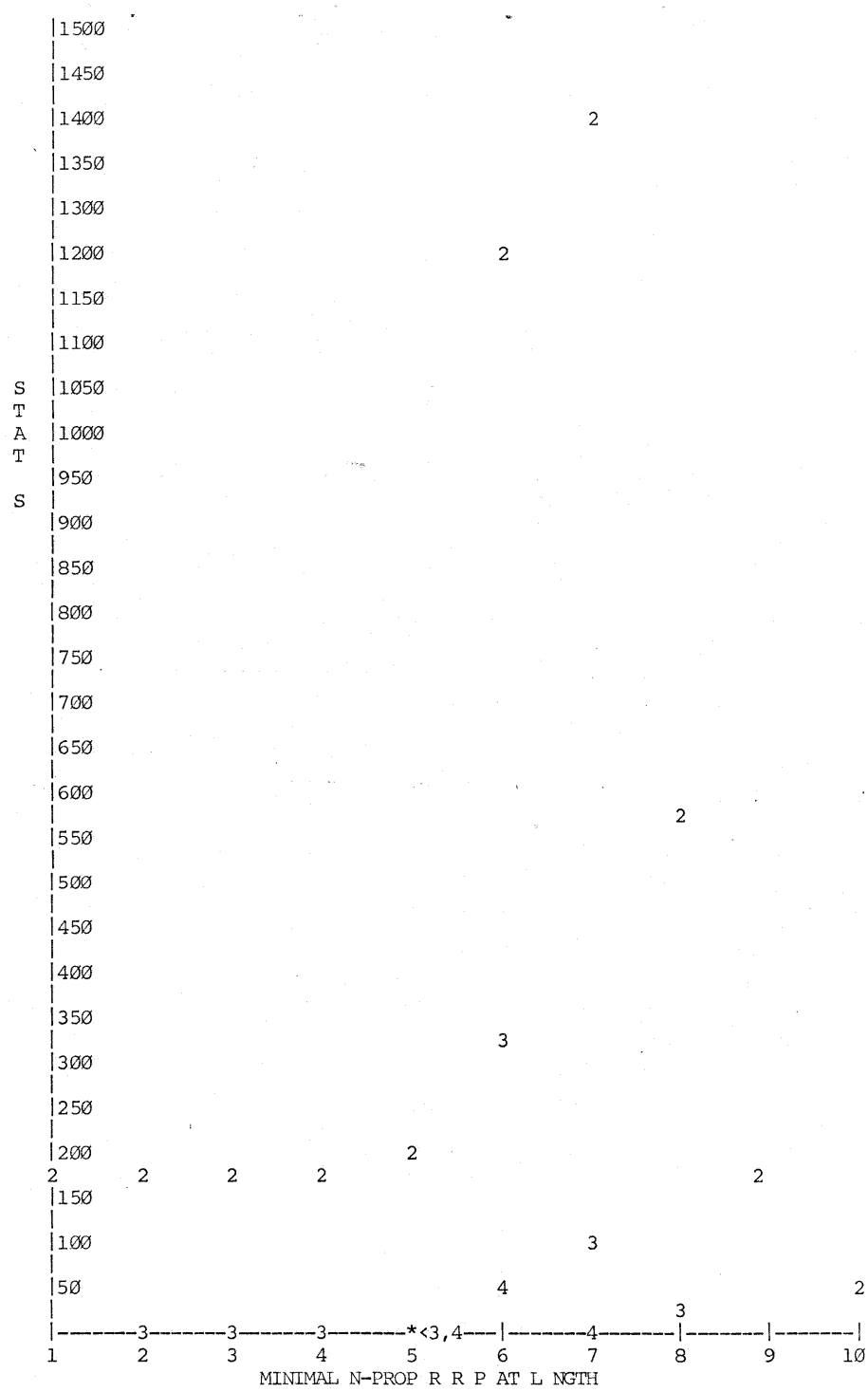
First of all, it is unclear whether DNA adequately satisfies the assumptions we have made. For example, of the portions of genetic material studied sufficiently to relate their functional properties to their internal structure, namely genes, the nucleotides seem to be functionally grouped in blocks of three; that is, from the initiation site of a particular gene the nucleotides seem to function in consecutive disjoint blocks where each block (called a codon) is three nucleotides in length. However, some genes overlap one another such that their respective codons sequences are shifted by one or two nucleotides; thus, these overlapping sections can be "read" two ways. Furthermore, there is an ample supply of intergene material whose function is unknown and may likely not be grouped functionally as codons.

In addition, the context filtering we have applied by increasing "n" to obtain more meaningful proper repeats will yield nothing if n is greater than 4 since this is the size of the alphabet. Therefore such low levels of analysis maybe difficult to interprete whenever there is a small alphabet. If, for example, when the number of states obtained is graphed as we did for English texts above, the difference between a DNA text and it's randomized version appear much more similar. To illustrate this consider the following graphs which are for the virus known as "fd" and it's randomized counterpart.



Graph of the states as a function of the minimal n-proper repeat length for genetic code of the virus fd.

2 denotes states for 2-proper repeats;
3 denotes states for 3-proper repeats;
4 denotes states for 4-proper repeats.



2 denotes states for 2-proper repeats
3 denotes states for 3-proper repeats
4 denotes states for 4-proper repeats

As a final comment on these graphs note that although they are fairly similar, this does not negate our earlier analysis concerning the existence and measurement of units. Since a significant number of units are 3 symbols long, a peak on the first graph at 7 is not inconsistent.

The above difficulties notwithstanding, when provided with the entire genetic code of simple organisms as a bacteriophage or virus, the program found sets of strings almost entirely 6 in length and such that a disproportionate number of them have their left boundary between the first and second nucleotide in the condons. It is worthwhile to remark that extremely restrictive parameters were used in finding these strings (4-proper repeats of minimal length 6) which if given as the parameters for an English text would yield a high a percentage of meaningful strings (assuming the text were long enough).

As a final comment note that at least some of the difficulties involving DNA that were discussed above could be circumvented by implementing an option to encode codons to represent the primitive input symbols for the program thus applying the program at a higher level of abstraction.

ACKNOWLEDGMENT

I gratefully acknowledge the assistance of Professor Andrzej Ehrenfeucht who provided much of the conceptual framework for this manuscript.

REFERENCES

[1] Aho, Hopcuoft, and Ullman, Analysis of Algorithms, 1976, Addison-Wesley, Reading, Mass., pg.347-357.

[2] Fries, C.C., The Structure of English, 1952, Harcourt, Bruce, and Company, New York.

[3] Wiener, Peter, "Linear Pattern Matching Algorithms", 1973, Conference Record, IEEE 14th Annual Symposium on Switching and Automata Theory, pg. 1-11.

[4] Harris, Zellig, "Discourse of Analysis", Language, 1950, vol. 28, pg. 1-50.

APPENDIX 1

EXAMPLE OF PROGRAM OUTPUT

Input text:

The house was used only for vacations, and vacations were not frequent.
However an older house such as this could not tolerate any more frequent
occupancy since it had not been kept in top condition. Of course top
condition had never been that great anyway.

minimal proper repeat length=5; size of "n" for n-proper repeats=2
number of states determined by non-nil n-proper repeats=16
text length=257; number of gaps having no n-proper repeats=223

(EXAMPLE OF OPTION 1: Note that the state numbers are not consecutive.
The integer value given a state is the index of the first gap contained
in that state.)

```
                    text                          textlength

 T h e  h o u s e  w a s  u s e d  o n l y  f    25
       ^          ^
       4          11
 o r  v a c a t i o n s ,  a n d  v a c a t i o    50
    ^             ^           ^
    28            38          28
 n s  w e r e  n o t  f r e q u e n t .  H o w    75
    ^     ^     ^ ^             ^
    38    58    62 63           71
 e v e r  a n  o l d e r  h o u s e  s u c h     100
^        ^             ^          ^
76       81            4          11
 a s  t h i s  c o u l d  n o t  t o l e r a t    125
                     ^ ^       ^
                   113 58      119
 e  a n y  m o r e  f r e q u e n t  o c c u p    150
              ^                  ^
              62                 71
 a n c y  s i n c e  i t  h a d  n o t  b e e    175
                    ^     ^^   ^   ^ ^
                    164 113 58 76 172 119
 n  k e p t  i n  t o p  c o n d i t i o n .     200
   ^          ^                        ^
   178        185                      164
```

```
  O f   c o u r s e   t o p   c o n d i t i o n   h   225
                ^                             ^
              185                           164
  a d   n e v e r   b e e n   t h a t   g r e a t   250
        ^       ^ ^       ^
       76    172 81     178
  a n y w a y .                                     257
```

(EXAMPLE OF OPTION 2:)

```
 proper repeats:
 gap:         4 length:         7
{ house }
 (     4,    11)
 gap:        28 length:        10
{ vacations}
 (    28,    38)
 gap:       114 length:         5
{ not }
 (    58,   119)
 (    58,    63)
 gap:        62 length:         9
{ frequent}
 (    62,    71)
 gap:        76 length:         5
{ever }
 (    76,    81)
 gap:       113 length:         6
{d not }
 (   113,   119)
 gap:       164 length:         6
{ had n}
 (   164,    76)
 gap:       172 length:         6
{ been }
 (   172,   178)
 gap:       185 length:        14
{ top condition}
 (   185,   164)
```

(EXAMPLE OF OPTION 3:)

```
 proper repeats listed by adjacent state numbers pairs:
 adjacent state numbers:       4,      11
{ house }
 adjacent state numbers:      28,      38
{ vacations}
 adjacent state numbers:      58,     119
{ not }
 adjacent state numbers:      62,      71
{ frequent}
 adjacent state numbers:      76,      81
```

{ever }
 adjacent state numbers: 113, 119
{d not }
 adjacent state numbers: 164, 76
{ had n}
 adjacent state numbers: 172, 178
{ been }
 adjacent state numbers: 185, 164
{ top condition}
 adjacent state numbers: 58, 63
{ not }

APPENDIX 2

PROGRAM OUTPUT FROM A LENGTHY TEXT

Here we present first a comparatively long text (16Ø61 characters) interspersed with designations of the state transitions when the text was processed for 3-proper repeats of minimal length= 8; secondly, we give a list of the proper repeats found.

To increase readability, the state transitions in the following text are denoted by the insertion of a "\" instead of the integer state designations. It should be noted, however, that 24Ø distinct states were obtained with 1284 state transitions. In this particular text, 86% of these state transitions are between words while the remaining transitions may be split into those that could be considered to have either a grammatical interpretation (e.g. "see\ing" or "sector\s") or those that break a word at a morphem (e.g. "electric\ity" or "produc\tion"), and, those that are exceptionally problematic (e.g. "i\n" or "la\wsuits").

it is important to understand the types\ of\ pollution \most important\ in\ japan\ a\nd to identify the origin\s\ of the \ \p\ollution \thus see\ing the\ \sector\s\ of the \economy \most effected by\ pollution\ \control\. the main sources\ of \ pollution\ \have been\ \industry \(extraction, process\ing and \fabrication)\,\ transporta\tion \and \final \ consumption\ \(waste \ production\).
\ industry\ \has been \forced by economical turns, the law and\ public \pressure to seek ways to produce that pollute less and\ are more \efficient. they must also produce goods \ with the\se characteristics. it\ has been \suggested that \ industrial \expenditures on\ pollution\ \control\ \increase\d \rapidly after 1968. by 1975, 18% of capital expenditures for \ industry \were on\ pollution\ \control\ \effort\s. this \figure is higher than in any other\ industrial \na\tion.\ the \ \japanese \have incorporated what they call the polluter pays principle (ppp) which forces the polluter to pay for \ pollution \either in fines or charges for not meeting standards (not really a socially acceptable solution any longer) or by installing\ devices \to prevent pollu\tion. the \polluter then incorporates the cost of prevention into the price of good\s. this \method is considered a more acceptable means of distribut\ing the\ \p\ollution \costs\ because \the consumer the pays\ for the\ \control\ \rather than the taxpayer\. government \support has assisted in encouraging \ control\s\ in the \form of grants, subsidies and loans for new equipment, tax incentives, contracts, and license approvals\. government \also\ provide\s administrative guidance\ with a \touch of paternal protection for business, but tries to

avoid action as an adversary or overriding authority.\ the most\ \recent\ \pollution \abatement\ devic\es \include \flota\tion and \separation\ devices \for water re-use\ treatment\ systems\, hooding\ systems\ for coke ovens, fugitive emission\ control\ \devices \and\ advance\d reverse osmosis for wastewater\ treatment\. thus\ pollution\ \control\, being of national interest, is becoming good\ for the \company's im\age and \a\ natural \part of\ production\ \responsibilities. it is important to note that at one point, many economists were \ concern\ed that excessive expenditures on\ pollution\ \control \ \might drive prices up and depress the economy. now, many argue\ that the\ \increase\d\ \demand for\ \research \and\ development\ \project\s and technology offset the decreased \ demand for \final goods. so, it is suggested that equilibrium will result\ with a \cleaner\ environment\, more satisfied \ public \and a better balanced economy.

solutions to th\e\ problem\ \of\ pollution \caused\ by the \ \public\ \are more \complica\ted and\ \problem\ \specific. plann\ing and\ \policy \implementation\ by\ the \government \can relieve many\ effects \but since there is no one particular offender, the responsibility and cost\s must be \shared by all. \ pollution \from the\ use of\ \public \or\ private \ \transportation \can be most effectively dealt with\ by the \manufacturer. auto manufacturers are producing more efficient cars that pollute less. their incentives are \ greater \cost effectiveness in\ production\, reduced\ use of \ \natural\ \resources \and\ energy \policies requiring them to do so.\ use of \mass\ transportation\ \in japan\ has always been high, but as the system improves schedul\ing and\ \people \realize the logic of using a les\s\ energy \intensive form of \ transporta\tion\ \and \the\re\ are more \vehicles for urban and suburban travel, more\ people \are \riding. again the manufacturer producing less polluting vehicles (\electric \ trains and busses)\ is the \root\ of\ pollution \abatement. other factor\s such as \traffic management and enforcement of noise and congestion laws alleviate som\e\ of the\ \p\ublic \ \problem\s.

the\ produc\tion\ \and \accumul\a\tion of \industrial,\ sew\age \and\ \solid waste \is a major\ concern\ for the \japanese\. japan\has few\ alternatives \for getting rid of waste. massive clean-up\ efforts \and new\ technologies\ \have been \provid\ing \ the \most \satisfactory results.

\ industrial \waste as a whole is difficult to discuss, but several factors can demonstrate\ recent \t\rends. plans for\ material\s recycling, extrac\tion of \reusable\ material\s from garb\age and \waste water and minerals recovery in mining operations are\ material\izing\. government \is subsidizing many\ of these \types of\ technologies \making them more attractive to industry. naturally, the\ effects \of waste disposal are\ significant \to\ economic\ \develop\ment\ \in \the \future\ because \a large por\tion of \waste created must be disposed of\ in\ japan\'s\ \limited\ \natural \setting\s. this\ \effects \the immediate areas\ and the\ \value of

surrounding\ resources \by taking away from\ japan's \already \ limited \reserves.
construc\tion of \urban\ sewage \facilities has only been promoted since 1958 when the\ sewage \law was enacted. by this time, population\ growth\ \in the \cities was already \ significant\ \and th\e\ \problem\ \soon\ became \overwhelming. the living\ environ\ment\ in \urban areas\ became \intolerable during this time\ because\ \pollution\ \control\ \was lax\ and the\ \construc\tion of\ \sewage \facilities was slow. more\ recent\ \efforts \have brought the invest\ment in \facilities to 1,669 billion yen in 1977 so that most cities over 100,0900\ people \are \served by it. in 1979,\ the \ japanese \goal was to\ increase\ \coverage of urban areas by 55% by 1985.\ recent \t\echnological emphasis\ has been \on tertiary\ treatment\, recycl\ing and \turning waste into compost and soil conditioner. there have also been of late regular international conferences to facilitate technological exchange on\ advance\ments.
\ solid waste \and\ dumping\ \have been \an eternal\ problem \ \in japan\.\ \limited \land and mass accumul\a\tion of \solid\ industrial \and\ \private \waste\ is the \root \ of th\e\ \p\roblem\. \ in the \smaller communities (less than 300,000), intensive group work to recover\ material\s has become popular. without using sophisticated or mechanized \ systems\,\ people \are \able to eliminate or reduce \ dump\ing \and \incineration. in organized groups,\ people \hand pick reusables or use screens or magnets to separate valuables, as appropriate. while this\ effort \doesn't make a huge dent\ in the \country's\ problem\, it represents a change in ethical view and a new attitude seeking self-sufficiency and possibly less\ consumption\.
urban\ solid wast\e\ \problem\s are far more complex and extensive\. government \is encouraging recycl\ing and \is financing various\ program\s \such as \mass compac\tion and \landfill\ project\s. to date, the situation is not improving drastically due to\ limited \choices for disposal but extensive\ research \will hopefully\ provide\ \answers soon. \ pollu\tion \and \its\ control\ \have had a\ recent \effect of limit\ing the \p\ath chosen\ by\ the\ \japanese \for\ economic \ \development\. as mentioned, until the mid-1960's\ economic \prosperity via\ growth\ \has been \the target of\ economic \ \policy\ \in japan\. \now, the value of continued expansion \ of the \economy is questioned as awareness of\ environment\al \limits force reconsideration. and\ pollution\ \problem\s are only one set of\ environment\al \limits forcing japan to seek \ alternative \means for maintaining their\ economic \stability. scarcity of\ natural\ \resources \has motivated \ the\ japanese \to find\ alternatives \to \dependence on non-renewable and foreig\n\ energy\ \resources \and to seek ways of extend\ing the \p\roductivity of their domestically available\ resources.\ \japanese\ \demand for\ \energy \has until recently been increasing faster than any western country. she is now second\ in the \world in absolute

\ consump\tion\ \of \primary energy, though japan does have the lowest per capita\ consumption\ \among major\ industrial \countries. the distribu\tion of \this\ consumption\ \between \ sector\s\ of the \economy \is quite different than most world patterns with\ industry \consuming 57.9% of raw energy\, \ transportation \only 18.7% and commercial and residential sectors the remaining 23.4%.\ increase\s i\n\ energy \demands \ have been \heaviest in\ \electric\al power and finished goods and\ have been \evenly distributed between personal use\,\ transporta\tion \and \industry.

\ with the\ past\ growth \of consumerism and emphasis of \ greater\ \production\, all\ sector\s\ of the \economy \ \have been \increasing their demands on\ natural\ \resources. \for instance\,\ transportation \needs are 4.5 times what they were in 1955. the annual\ demand for \non-renewable \ resources\ \has been \escalating since the 195Ø's. iron, aluminum\ and other \minerals are becoming scarce as mining operations expand throughout the country. there\ has been \an\ increase\ \in the\ \demand for \wood, but more\ significant \ \is the\ \increase\ \in desire for animal protein among the people. this means a\ greater\ \need for \grain and fodder crops as well as land and equipment to produce more animals. a\ recent\ \concer\n\ in the \ \industrial \and \household sector\ is the\ \growing\ \need for \fresh water\ resources \ \in the \face of diminishing availability. the\ effects \of\ increase\d\ \consump\tion\ \of \oil are also\ significant \since 99.8%\ of the \oil japan needs is impor\ted and \oil accounts for 88% of it\s\ energy \requirements. though\ japan is \advancing rapidly as an international leader, the idea of continued\ economic\ \development\ \a\nd \ consump\tion\ \of\ \resources \at high rates is\ limi\ted \and \is prompting\ research \into\ alternatives \to \conventional \ resource \supplies. these\ alternatives \will be examined shortly.

it\ has been \suggested that\ environment\alism as a national \ concern\ is\ growing\ \in japan\. \seeing\ japan's\ \recent \history and viewing its\ economic \turns have hopefully presented a basis for understanding som\e\ of the\ \p\ollu\tion \and\ \resourc\e\ \problem\s.\ the\ japanese \are looking for avenues towards their continued\ development\ \with some limitation, but in balance\ with the\ir few\ natural \ \resources\.

\government \seeks this goal\ through\ \policy \for business, \ industrial \and \domestic sectors, business and\ industry \pursue technological\ advance\ and the\ \people \are \becoming mobile in their\ effort \to communicate a\ need for \new values via a strong\ environment\al \movement.

major protests bega\n\ in the \mid-196Ø's challeng\ing the \foundations and values upon which\ japanese\ \economic \ \growth \began. in early history incident\s such as \the ashio incident where the hitachi mining company's refinery emitted sulfurous acid gasses in such quantity \ that the \surrounding forests were killed. conditions

\ in the \area were so deplorable that many workers were unable to continue to work. the victims in this case exerted sufficient pressure on the company to cause the company to construct a high chimney, allow\ing the \p\revailing winds to carry the toxic gasses to the sea. the success of similar\ efforts \was relatively slight until after the new constitution which gave freedom of press, speech and organiza\tion. the \organizations of\ recent \t\imes are spontaneous, local and mostly volunteer. most oppose\ development\, work for health compensation or encourage\ alternative \ways of living. the strategies used\ by the \groups include negotiation\ with a \polluter, distribu\tion of \information, mass demonstration\s and la\wsuits.\ the most \publicized groups are the society for non-wasteful \ consumption\, kansai recycling movement and various anti-highway and anti-construction groups. most\ of these \groups are regional at the largest scale; small but active\ is the \motto for\ environment\al \groups\ in japan\. \now in order to complete the spectrum of analysis on the situation surrounding\ resource \scarcity and\ pollution \ \problem\s\ in japan\, a short examin\a\tion of \potential new\ energy\ \resources\ \through\ \research \follows. the basis of\ japanese\ \resource \and\ pollution\ \problem\s is\ growing\ \consump\tion\ \of \energy;\ energy \of all form\s must be \derived from\ natural\ \resource\s \and la\rge investments in them\ provide\ \for the\ \growing \needs of a\ growing \popula\tion.

\ the\ \japanese\ \have been \exploring\ alternatives \for \ energy\ \pro\duction\ \for many decades, but it was not until 198Ø\ that\ the \government \gave serious attention to shifting\ japan'\s\ \energy \base. 198Ø was designated as "alternative energy: the first year" and gave\ development \ \of\ alternativ\e\ \energy \high budgetary priority. in october of that year,\ the government \also established the new\ energy\ \development \ \a\uthority to explore and implement\ alternativ\e\ \energy \pro\grams.\ japan's \main priority is to diminish dependence on foreign oil\ resources \(75% down to 5Ø% by 199Ø). they expect to reach this target\ through \diversific\a\tion of \suppliers,\ increase\d\ \development\ \on foreign soils, \ energy \conservation, and\ development\ \of\ alternativ\e \ \energy\ \resources.

\the major thrust of this last\ effort \started in 1974 \ with the\ sunshine\ project\ which leads\ research \in solar, thermal,\ coal and \hydroge\n\ energy\ \pro\duc\tion\. the \moonlight\ project\ promotes\ research \and\ development \ \of\ energy \conservation techniques in an\ effort \to mak\e\ energy \goal\s\ of the \future quickly attainable. \ japanes\e\ \energy\ \policy \suggests that\ nuclear \power can contribute\ the most \to it\s\ energy \supply system. plants\ provide\d 1Ø.8% of\ energy \needs in 198Ø\ and \ the\ \government \plans to\ increase\ \dependence to 16.7% by 1985. their major\ concern\s in\ nuclear\ \development

\ \a\re safety, waste disposal\ and the\ \lack of uranium
\ in japan\. \they have responded to the first\ problem
\ \by reorganiz\ing the \atomic\ energy \commission to
establish a separate\ nuclear \safety commission. with
no appreciable amount of uranium,\ the\ japanese \support
\ programs \that make\ the most \of existing supplie\s such
as \fast breeder reactors and\ advance\d thermal reactors.
\ nuclear \fusion is being explored\ becaus\e\ \of the
\inexhaustible supply of hydrogen found in water to fuel the
reactors. heat\ pollution \is a\ problem\ \with thi\s\ energy
\ \resource \and spent fuels\ and other \radioactive residual\s
must be \perfectly contained\.\ japan \is \experimenting with
extensive sea\ dumping\ \program\s \and la\nd isolation
for toxic\ material\. also a major program for storage
facilities in underway.
coal powered plants, while\ in the \past\ became
\unattractive and uneconomical are reawakening for a
transitionary\ \electric\ity\ supply. new ways to utilize
\ coal and \new ways to use the tremendous amounts of ash
produced are among the newest\ project\s.\ coal and \oil
mixtures and coal gasification are two\ technologies
\being explored.
liquid\ natural \gas exploration in alaska\ and other
\part\s\ of the \word for supplies to\ japan a\re becoming
more feasible and\ private\ \electric\ power companies
are beginning to make the large scale investments necessary
to make the\ resource \available. hydro\electric\ power
comprised 4/5 of all\ \electric\ power until 1955. thermal
and\ nuclear\ \development\ \now lead hydro power, but it
could spring back as all\ alternatives \are being pursued.
\ the\ japanese \are also on\e\ of the \s\trongest proponents
internationally of renewabl\e\ energy\ \resource \research. as
suggested by implement\a\tion\ of \the \s\unshine\ project\,
japan places emphasis on applic\a\tion of \technology that
fits the need\s\ of the \s\ociety. many\ of the \applications
are of a 2Ø to 3Ø year plan\ and the\ir contribution may
not be evidenced until the year 2ØØØ.
large scal\e solar \utilization as well as small\ private
\applications have had\ advance\d breakthroughs in\ recent
\years. power towers for generating\ \electric\ity\, and
solar furnaces are on the drawing board or under construction.
small scale\ technologi\es \include \th\e solar \cooled
home and progressiv\e solar \hot water heating\ systems\.
geothermal\ energy \(subterranean heat turned into
\ \electric\ity\) has terrific potential\ in japan\, on\e\
of the \world's most volcanic areas. hot springs area are
reached by drilling wells and permitting steam to operate
power generators. wave power, ocean currents and temperature
gradients all have tremendous potential in generating power\.
\ japan \a\lso seeks to cut\ consumptio\n\ \in the\ \private
\sector\ through \conservation. measur\es include \closing gas
stations, restricting late television broadcasting, fewer
neon signs with shorter lighting hours and fewer late

business hours. no car mondays, earlier baseball games and lower speeds for ships and cars shall also contribute to \ energy \savings-16. all\ of these\ \efforts \separately and combined can\ provide\ \lasting solutions to\ japan'\s \ \energy\ \resource \scarcity.

in sum, with all parts of society pursuing many answers and various\ alternatives \to \the difficulties of\ growth \in today's world,\ japan is \rapidly putting herself in a more self-sufficient and prepared position to deal with futur\e\ energy\ \pro\blem\s.

List of proper repeats: " advance", "age and ", " alternative ", " alternatives ", " alternatives to ", " and other ", " and the", " and the ", " are more ", "ation of ", " became ", " because ", "by the ", " coal and ", " concern", " consumption", " consumption ", " consumption of ", " control", " control ", " demand for ", " development", " development ", " development a", " devices ", " dumping ", "e energy", "e problem", "e of the ", "e solar ", " economic ", " effects ", " effort ", "electric", " electric", " electricity", " energy ", " energy pro", " environment", " environmental ", "es include ", " for the ", " government ", " greater ", " growing ", " growth ", " has been ", " have been ", " in japan", " in japan. ", " in the ", " increase", " increase " , " increased ", " industrial ", " industrial and ", " industry ", "ing and ", "ing the ", " is the ", " japan a ", " japan is ", " japanese ", " japan's ", " limited ", " material", "ment in ", " natural ", "n energy ", " need for ", "n in the ", " nuclear ", " of pollution ", " of the ", " of the p", " of the s", " of these ", " people ", " people are ", " policy ", " pollution ", " pollution control", " private ", " problem", " problem ", " production", " production ", " programs ", " project", " provide", " provide ", " public ", " recent ", " recent t", " research ", " resource ", " resources ", " resources. ", " research ", "s and la", "s energy ", "s must be ", "s such as", "s of the ", "sectors of the economy ", " sewage ", " significant ", " solid waste ", " systems", "s. this ", " technologies ", "ted and ", " that the ", " the government ", " the japanese ", " the most ", " through ", "tion and ", "tion of ", "tion. the", " transportation ", " treatment", " use of ", "with a ", " with the", ", transportation ", ". japan ".

%

APP NDIX 3

CHILDR N'S T XT US D AS PROGRAM INPUT

The text below is sufficiently long such that it is has been reprinted here with the state transitions marked as the text in Appendix 2 is. Note that the parameters used are also the same as for this other text; i.e. 3-proper repeats of minimal length 8.

Red Man, White Man, African Chief The Story of Skin Color
by Marguerite Rush Lerner, M.D.
Lerner Publications Company, Minneapolis, Minnesota

all\ living things \have color: plants, animals, people.
the stuff that color is made of is\ called\ \pigment\.
\paints are a form of\ pigment\. \the yellow\ color\ of
\the \buttercup\ comes from the\ pigment\, \xanthophyll.
dandelions\ contain\ \xanthophyll, and so do the leaves
that turn yellow in autumn. the orange\ color\ of \the
\carrot\ comes from the\ pigment\, \carotene. pumpkins
\ contain\ \carotene. the green\ color of \grass\ comes
from the\ pigment\, \chlorophyll. asparagus, spinach
and lettuce have chlorophyll. the red\ color of \blood
\ comes from the\ pigment\, \hemoglobin. hemoglobin
\ contain\s the metal, iron. some rocks look red\ because
\they\ \contain\ \iron, too. what is black? charcoal is
black. the chemical, carbon, makes charcoal black.
blackness of\ living things \is not due to carbon. a
\ pigment\ \called \melanin makes living cells black or
brown. if you cut open a raw potato and let it stand
awhile, it turns brown or black. that happens\ becaus\e
\ \melanin\ pigment\ \forms when an uncooked potato is
exposed to air. a banana becomes brown or black when it
is cut open and left out\ in the \air or when it is
bruised.\ melanin\ pigment\ \forms\ in the \banana.
mushrooms can make lots of melanin.\ living things
\are made of different kinds of cells.
each cell gas a job to do. this is a\ melanocyte\. a
\ melanocyte\ is a cell that makes\ melanin\ pigment\ \or
coloring matter. the zebra's stripes and the leopard's
spots are made of melanin. the skin\ of the \frog is
spattered with bunches of\ melanocyte\s that look like
ink blots\.\ people \have\ \melanocyte\s, too,\ in\ their
skin\, hair roots and eyes. some\ people \have \mor\e
\ melanin\ pigment\ \than others. many negroes have a lot
of\ melanin\ pigment\. \some\ people \have \very little
or no\ melanin\ pigment\. \they look white and are\
called \albinos. other\ people \have \in-between amounts.
the american indian does not have red skin.\ his skin
\is brown\ because \it\ contain\s\ melanin\ pigment\.
\a person is a "red man" only if he paints\ his skin

\red\.\ people \from china, japan and india\ do not \have yellow skin.\ their skin\ is brown\ because \it\ contain\s \ melanin\ pigment\. \the white man is not really white \ because\ \his skin \does\ contain \melanin\ pigment\. \when he goes out\ in\ the \sun\,\ his skin \becomes brown\ because\ \melanocyte\s make more\ pigment\ \during hot weather with plenty of sunshine\.\ people \who sit by an open fire or who use a heating pad for a very long time may get brown spots on\ their skin\. freckles are dark spots\ in the \skin that\ contain\ \melanin. they first form in children about 5 years old after they have been out\ in\ the \sun\. freckles are darker, and there are more of them,\ in the \sumer than\ in the \winter. they become less noticeable after a person reaches the age of 25. moles or nevi are dark spots\ in the \skin that\ contain \ \melanin. they\ do not \need\ sunlight\ to form. nevi first show up in children about 3 years old\.\ people \get more as they grow older. all\ people \have \about the same number of\ melanocyte\s\ in\ their skin\. albinos have milanocytes, but\ their skin\ is white\ because \they \are born without a special chemical\ called \tyrosinase that is needed to form\ melanin\ pigment\. \red-headed\ people \ \do not \suntan as well as\ people \with \blond or dark hair\ because \they\ \do not \have enough tyrosinase needed to make\ pigment\ \in\ their skin\. red-haired persons tend to sunburn more than to suntan.\ dark skin \helps protec\t\ people \from strong\ sunlight\. nobody knows whether the first person on earth had\ dark skin \or \ light skin\. we do know that in parts\ of the \world where there is a lot of\ sunlight\, as in africa and india, mos\t\ people \have \dark skin. where there is less sun and heat, as in northern europe and asia and north america, mos\t\ people \have\ \light skin\. many thousands of years ago, man lived outside more than he does now. he did not have a house to keep out\ the sun\shine. it is possible that\ in the \hot countries\ people \with\ \dark skin \were able to survive\ because \th\e\ melanin\ pigment \ \in\ their skin\ protected them against the bright rays \ of\ the \sun\. in northern countries with less\ sunligh\t \ people \did not need as much\ pigment\ \to stay alive. the skin\ color of \our ancestors who lived thousandes of years ago has been passed down to us who live today. usually\ people \with\ \light skin\ have children with \ light skin\, and\ people \with\ \dark skin \have children with dark skin. the skin that covers our bodies is like cloth. cloth is woven from threads of many colors. the\ color of \our skin comes from the different\ pigment\s that we get from our parents and from our ancestors of long ago.