Experiments were conducted on a book, Current Medical Information and Terminology, (AMA, Chicago, 1971, edited by Burgess Gordon, M.D.), which is a compendium of 1262 diseases, each of which is defined by a collection of attributes. The original purpose of the book was to introduce a standard nomenclature of disease names, and the attributes are organized in conventional medical form: a definition consists of a brief description of the relevant symptoms, signs, laboratory findings, and the like. Each disease is, in addition, assigned to one (or at most two) of eleven disease categories which enumerate physiological systems (skin, respiratory, cardiovascular, etc.). While the editorial style of the book is highly telegraphic, with many attributes being expressed as single words, it is nevertheless easily readable (see Figure 1).

The vocabulary employed consists of about 19,000 distinct "words" (determined by a lexical definition), roughly divided equally between common English words and medical terms. We measured word frequency by "disease occurrence", (the number of disease definitions in which a given word occurs one or more times). By this measure, only seven words occurred in more than half the disease definitions, and about 40% of the vocabulary occurred in only a single disease definition. (Table 1 lists the words at the top of the frequency list together with the number of occurrences.)

Assisted by the facilities of the TMuNIX operating system, we created a series of inverted files (from a magnetic tape of the CMIT text), and developed a set of interactive programs to form a word-and-context query system. This system has enabled us to study the problem of inferring term reference in this large sample of text (some 133,000 word occurrences), within the context of diseases.

An interesting early result was the ease with which many medical terms could be algorithmically separated from common English words. After adjusting for the fact that some disease categories are larger than others, we defined an entropy-like measure of the distribution of word occurrences over the eleven physiological categories as a measure of category specificity. We reasoned that some medical terms such as 'murmur', while not specific to any particular heart disease, are specific to heart disease generally. This term would not, for example, be used in describing endocrine disorders. Such a word would be expected to occur in category 04 (cardiovascular disease) frequently, and not in the other categories. Such a term would, by our measure, have a low 'entropy'. A common English word like 'of', would be used in the descriptions of all kinds of disease, and would accordingly have a high 'entropy'.

Tables 2 and 3 show the top and bottom of the list of all words occurring in two or more diseases sorted by this entropy measure. In these lists, as our hypothesis seems to imply, low 'entropy' corresponds to high 'specificity', and high 'entropy' to low 'specificity'. This separation of medical terms from common English words, by algorithmic means, is facilitated by the context supplied by the notion of 'disease category', and the fact that this was represented in the CMIT text.

Our second experiment investigated the co-occurrence properties of some medical terms. Aware that many medical diagnostic programs have assumed attribute independence, we sought to shed light on the appropriateness of the assumption by evaluating it in terms of word co-occurrence in disease definitions.

Since the previously described procedure had given us a means of selecting medical terms from common English words, it was possible to produce lists of 'pure' medical terms. We then wrote a program which formed all pairs of such terms (ignoring order). We defined an 'association measure' (A) which measured the difference between the observed co-occurrences of term-pairs (they could co-occur in any location in the definition and in either order), and the co-occurrences expected from chance alone. Tables 4 and 5 show the top and bottom of a list of all pairs formed from the low entropy terms in the previous experiment. The first 1120 terms were chosen, that is, those having an entropy of 2.0 napier or less. The pair list was then sorted by this association measure, A.

Word pairs which are found to be highly associated, appear to do so for two reasons. The test, which is trivial, is that some word pairs are semantically related. Words despite their being lexically, two. Common examples would be 'White House' and 'Hong Kong'; medical examples are 'vital capacity', 'axis deviation', and 'slit lamp'. These could have been avoided algorithmically by not taking adjacent words in forming the term-pairs, without any significant overall effect. The second reasons for high frequency word co-occurrence is that both words are causally related through underlying physiological mechanisms. It is these which had the greatest interest for us, and the measure A, may be viewed as a measure of the non-independence of the symptoms or signs themselves.

The term pairs which are negatively associated, have this property for the same reason. If the two terms are used typically in the descriptions of different diseases, they are less likely to co-occur than by chance. (In a baseball story on the sports page, we would not find 'pass', 'pun', or 'tackle'). These negatively associated pairs may have value in diagnostic programs for the recognition of two or more diseases in a given patient, a problem not satisfactorily dealt with by even the most sophisticated of current programs.

Finally, an extension of the entropy concept permits one to generate (algorithmically) the vocabularies used by the medical specialties which correspond to the disease categories represented in CMIT. This is done by assigning terms which occur predominantly in one category to a single vocabulary and then sorting by entropy. Tables 6 and 7 show the vocabularies used in dermatology and gastroenterology (as derived from CMIT). These vocabularies, it will be noted, can be used as 'hit lists' for the purpose of recognizing the content of medical texts.

In summary, we see the ability to differentiate medical terms from common words by context, and the ability to relate the medical words by meaning, as two of the first steps toward text processing algorithms that preserve and can manipulate the semantic content of words in medical texts.
COLORADO TICK FEVER 00 2217
AT FEVER, MOUNTAIN; FEVER, MOUNTAIN TICK.
ST VIRUS TRANSMITTED BY TICK DEMECENTOR ANDERSONI.
SM CHILLS; HEADACHE; PHOTOPHOBIA; BACK-
ACHES; PAIN IN EYE; MYALGIA; ANOdynE;
NAUSEA; VOMITING; PROSTRAITION.
SG SEASONAL, MARCH TO JULY, IN WESTERN
UNITED STATES; INCUBATION PERIOD 4-6
DAYS; ONSET ABRupt; POSSIBLY SLIGHT
SICKNESS; SUSTAINED FEVER, 102-104 F
OR HIGHER SIGNIFICANT; PULSE RATE
INCREASED, COURSE IN PREVENTION,
REMozAL OF TICK FROM SKIN; APPLICaTIONS
TO SKIN OF TURPENTINE, IODINE,
ACETONE; REMOVAL OF TICK BY INSERTION
OF NEEDLE BETWEEN MOUTH PARTS; ASPIRATION
FOR PAIN; ANTIBIOTIC TREATMENT IN-
EFFECTIVE.
CM ENCEPHALITIS, MENINGITIS ESPECIALLY IN
CHILDREN.
LB WBC DECREASED; MONOCYTOSIS; COMPLIT-
MENT FIXATION TEST POSITIVE; INJECTION
OF SERUM OR CSF KILLING SUCCLING
MICKE; NEUTRALIZATION OF VIRUS WITH
IMMUNE SERUM RESULTING IN SURVIVAL.

Figure 1. Typical disease 'definition'
taken from CMIT

Table 1. The highest frequency words
used in CMIT, together with the number
of disease definitions in which the
word occurs at least once.

| Word | Frequency | Definitions |
|------|-----------|-------------|
| Fever | 250 | 123 |
| Mountain | 150 | 75 |
| Tick | 120 | 60 |
| Disease | 100 | 50 |
| Symptoms | 90 | 45 |
| Headache | 80 | 40 |
| Nausea | 70 | 35 |
| Vomiting | 60 | 30 |
| Prostration | 50 | 25 |
| Chills | 40 | 20 |
| Photophobia | 30 | 15 |
| Backache | 20 | 10 |
| Pain | 10 | 5 |

Table 2. The lowest 'entropy' words
in CMIT, in order of increasing 'entropy'
The entropy is given in the first column; the
entries in the next 11 columns are the
percent of occurrences in the 11
disease categories (body as a whole, skin,
musculo-skeletal, respiratory, cardiovas-
cular, hemic and lymphatic, GI, GU, en-
docrine, nervous, organs of special sense).

| Word | ENTROPY |
|------|---------|
| Fever | 0.01 |
| Mountain | 0.02 |
| Tick | 0.03 |
| Disease | 0.04 |
| Symptoms | 0.05 |
| Headache | 0.06 |
| Nausea | 0.07 |
| Vomiting | 0.08 |
| Prostration | 0.09 |
| Chills | 0.10 |
| Photophobia | 0.11 |
| Backache | 0.12 |
| Pain | 0.13 |

150
Table 3. The highest 'entropy' words in CMIT. Note that these are common English words.

| Word | Entropy | Word | Entropy | Word | Entropy |
|------|---------|------|---------|------|---------|
| and  | 0.9608  | the | 0.9590  | a | 0.9566 |
| of | 0.9598 | of | 0.9596 | be | 0.9584 |
| in | 0.9592 | in | 0.9590 | to | 0.9578 |
| that | 0.9586 | that | 0.9584 | is | 0.9562 |
| on | 0.9582 | on | 0.9580 | for | 0.9558 |
| with | 0.9578 | with | 0.9576 | by | 0.9554 |
| not | 0.9574 | not | 0.9572 | as | 0.9540 |
| for | 0.9570 | for | 0.9568 | of | 0.9536 |
| and | 0.9566 | and | 0.9564 | and | 0.9532 |

Table 4. The top of the word-pair list in decreasing order of association value (A).

Table 5. The bottom of the word-pair list, showing the negatively correlating words.
### Table 6. A word list generated algorithmically which constitutes a dermatological vocabulary. The disease category 'skin' is represented by the third column.

| Disease Category | Skin | Other | Digestive | Renal | Endocrine | Nervous | Musculoskeletal | Respiratory | Gastroenterology | Immunological |
|------------------|------|-------|-----------|-------|-----------|---------|----------------|------------|-----------------|--------------|
| Acne             |      |       |           |       |           |         |                |            |                 |              |
| Eczema           |      |       |           |       |           |         |                |            |                 |              |
| Urticaria        |      |       |           |       |           |         |                |            |                 |              |
| Dermatitis       |      |       |           |       |           |         |                |            |                 |              |

### Table 7. A word list generated algorithmically which constitutes a vocabulary of gastroenterology. The eighth column represents the disease category 'digestive system'.

| Disease Category | Digestive | Renal | Endocrine | Nervous | Musculoskeletal | Respiratory | Immunological | Gastroenterology | Immunological |
|------------------|-----------|-------|-----------|---------|----------------|-------------|--------------|-----------------|--------------|
| Gastritis        |           |       |           |         |                |             |              |                 |              |
| Hepatitis        |           |       |           |         |                |             |              |                 |              |
| Appendicitis     |           |       |           |         |                |             |              |                 |              |
| Diverticulitis   |           |       |           |         |                |             |              |                 |              |

---

**Table 6** and **Table 7** highlight the algorithmically generated word lists for dermatological and gastroenterological vocabularies, respectively, with various disease categories included.