124 SLACK, W.V., SLACK, C.W.: Good Questions and Bad in Anderson, J., Forsythe, J.M. (ed.): Information Processing of Medical Records, Amsterdam - London (1970) pp 133-143
125 SLACK, W.V.: Computer Based Interviewing System Dealing with Nonverbial Behaviour as well as Keyboard Responses, Science 171 (1971) 84-87
126 SLACK, W.V., SLACK, C.W.: Patient Computer Dialogue, New Eng.J.Med. 286 (1972) 1304-1309
127 STERLING, T.D., GLESER, M., HABERMAN, S., POLLACK, S.: Robot Data Screening: A Solution to Multivariate Type Problems in the Biological and Social Sciences, Comm.A.C.M. 9 (1966) 529-532
128 STERLING, T.D., HABERMAN, S., POLLACK, S.: Robot Data Screening, an Automated Search Technique, Biomed.Comp. 1 (1970) 61-73
129 VECCHIO, T.J.: Predictive Value of a Single Diagnostic Test in Unselected Populations, New Eng.J.Med. 274 (1966) 1171-1173
130 WAITZKIN, H., STOECKLE, H.: The Communication of Information About Illness - Clinical, Sociological and Methodological Considerations, Adv. Psychosom.Med. 8 (1972) 180-215
131 WARNER, H.R., OLMSTED, C.M., RUTHERFORD, B.D.: HELP - A Program for Medical Decision Making, Comp.Biomed.Res. 5 (1972) 65-74
132 WARNER, H.R., RUTHERFORD, B.D., HOUTCHENS, B.: A Sequential Bayesiaen Approach to History Taking and Diagnosis, Computers Biomed.Res. 5 (1972) 256-262
133 WARNER, H.R..: Health Evaluation Through Logical Processing - HELP Kansai Institute of Information Systems, Proc. Medis'73, Osaka (1973) pp 197-303
134 WARNER, H.R., MORGAN, J.D., PRYOR, T.A., CLARK, S., MILLER, W.: HELP - A Self-Improving System for Medical Decision Making, in Anderson, J., Forsythe, J.M. (ed.): Medinfo'74, Amsterdam: North Holland (1974) pp 989-993
135 WEED, L.L.: Medical Records, Medical Education and Patient Care. (Cleveland: Press of the Case Western Reserve University 1969)
136 WESTRIN, C.G.: The Reliability of Auto-Anamnesis - A Study of Statements Regarding Low Back Trouble, Scand.J.Med. 2 (1974) 23-25
137 WIDROW, B.: Generalization of Information Storage in Networks of Adaline "Neurons", Marshall et al. (ed.): Washington, (1962) pp 453-461
138 YOUDEN, J.: Index for Rating Diagnostic Tests, Cancer 3 (1950) 32-35
139 YARNALL, S.R., SAMUELSON, P., WAKEFIELD, B.S.: Clinical Evaluation of Automated Screening History, North West Medicine (1972) 186-191
140 YOUNG, D.W.: Assessment of Questions and Questionnaires, Meth.Infor.Med. 10 (1971) 222-228
141 YOUNG, D.W.: Comparison of Information Collected by a Questionary with that in the Patient's Hospital Record, Meth.Inf.Med. 11 (1972) 20-22
142 YOUNG, D.W.: Evaluation of Questionnaire, Meth.Inf.Med. 11 (1972) 11-19
143 ZEINER-HENRIKSON, T.: The Repeatability at Interview of Symptoms of Angina and possible Infarction, J.Chron.Dis. 25 (1972) 407-414
144 ZERSSEN, D.V.: Die Beschwerdenliste als Test, Therapiewoche 21 (1971) 1908-1920.

# MEDICAL LANGUAGE DATA PROCESSING

Friedrich Wingert, Münster

## 1 Need and Application Areas for Formal Representation of Language Data

The present form of the medical record has been developed when patient care was primarily a single physician-patient relation. It has been used by the physician to record his observations and orders. Thus the medical record has been, in the legal sense, the rationale of clinical decisions and actions. It is most important that the data have been written down by a physician to be read again by himself, if necessary. This fact has led to an extremely physician-related document with subjective associations, omissions, ordering, nomenclature, style etc.. The largest part of the medical record contains free text, often handwritten and illegible to other people.

In modern medicine the environment has changed, but document has not. Several physicians, nurses and technical personnel take care of a patient in a hospital and several medical specialities are engaged in the process of diagnostics, therapy and surveillance [1]. All these people may add data to the medical record which suffer from the same weaknesses. The result is a fairly unstructured document, unreliable with respect to completeness and correctness and in general useless for research. Hundreds of medical students each year and not only students extract data from these documents for "research".

In this situation it is not sufficient to argue that the system is running. It is unknown, how many important data have not been captured, how many data have been captured several times, how often relevant data in the medical record are not recognized or are misunderstood by an attending physician, especially when he is short in time, and how many hours medical students waste their time in generating tables and false hypotheses or even "proofs" from medical records.

The goal is the generation of documents which really support health care delivery in every day routine and in research, i. e., communication, justification and analysis. Tasks that have to be done to reach this goal are

- *standardization* of the medical terminology, procedures and data structures,
- explicit *rules* for data aggregation and communication.

It seems to be reasonable to investigate, how far qualitative characteristics for computer programs, like portability, adaptability and reliability [10] can be applied to the medical record.

Basically there is only one alternative to natural language-based systems in the documentation of medical language data: Language data have to be coded by men into a highly formalized artificial language. Experience shows, that physicians are in general neither willing to learn such an artificial language and work as a code clerk, nor are they well performing this job. In other words, special personnel has to be used to translate natural language statements into artificial language statements. Despite the fact that such personnel is expensive and is available in a limited number only, the results are loaded with a high error rate due to errors in interpretation and transcription and due to inconsistency of interpretation.

Because most of the data in medicine are in language form, use of computers in medical data processing directly leads to the problems connected with the processing of language data. If computers are accepted as a powerful tool for the improvement of communication of data and information, than the problems of processing of language data have to be solved.

Communication in general is the application area which can be subdivided according to the links of communication chains:

| | |
|---|---|
| Physician A ⟷ file ⟷ Physician A, B: | Report, consultation, document, medical record. |
| Patient ⟷ Physician: | Diagnostic chain (Speech analysis: drug influence, psychoses); medical history taking; orders. |

Another scheme can be derived from clinical routine:

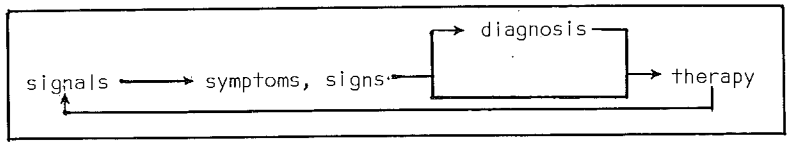
- Initiating and performing the diagnostic and therapeutic procedures,
- timely collecting, summarizing and reporting data,
- storage and retrieval of medical records for retrospective studies.

In the broader sense, language used in medicine is natural language, which becomes more specific when only physicians or paramedical personnel are linked in the communication chain.



## 2 Goals of Automated Processing of Medical Language Data

Let me first give a rough scheme of the decision process in clinical medicine:




The signals are perceived by physicians and are interpreted as ***symptoms*** or ***signs***. Symptoms and signs are collected and can be abstracted and classified as a ***diagnosis***. Symptoms and signs either lead directly or via diagnosis to a ***therapy*** which may generate new signals. These processes include comprehensing and generating language-based informations and indexing of these informations due to accepted classification schemes.

Computer-based processing of medical language data involves

- *recognition* resp. *analysis* (morphological, syntactical, semantical) of a language string,
- *representation* of the content using a data structure.

### 2.1 FUNCTIONAL GOALS

Functional goals are derived from the mental processes involved in extracting, collecting, ordering and retrieving informations in medicine. They are:

- Automated *indexing* of medical utterances due to a given classification scheme.
- Automated *abstracting* of medical documents.
- Automated *retrieval* and counting of medical facts and documents.

There is a significant gap between theoretical terminology and language which is used in practice. This gap may lead to contradictions between several goals such as serving of local needs versus international compatibility.

### 2.2 DERIVED GOALS

Derived goals are combined functional goals and/or are combined with functional goals of areas of interest other than medical language data processing. Examples are:

- Support for medical decision making,

- support for medical education,
- generation of medical reports,
- machine translation,
- health cost estimation,
- linking of different classification schemes,
- generation of hypotheses about syndromes that have to be verified genetically, etiologically or pathogenetically,
- language-based question-answering systems.

## 3 Components of Language Data Processing

### 3.1 DEFINITIONS

Natural language: This is not the place and I do not have the competence to evaluate the extensive literature on natural language. So far, there seems to be no definition of natural language which is fundamentally different from: ***General-purpose tool of communication within a speech community***. That means, language is used by a transmitter to encode informations out of a human information store and has to be decoded by a receiver. Language elements are words, signs and rules for combination of words in such a way, that a message can be understood by human beings in a special domaine of discourse.

Natural language in medicine: Fuzzy subset of natural language which is used by competent speekers in medicine speeking about medicine.

Artificial language: Designed for a restrictive purpose by and for specialists for use in a limited domaine of knowledge and application [25].

An artificial language consists of words and rules for combining words. Mapping of an utterance formulated in natural language into an utterance formulated in an artificial language has to be done in such a way that the information content of both formulations is approximately identical.

Medical language: Artificial language for the communication of medical informations.

Medical language is not just natural language in medicine. As we will see later, there are additional rules and there is dominance of semantics over syntax which is different from natural language. On the other hand, medical language lacks of most of the advantages present in highly formalized artificial languages, like (1) precision, (2) axiomatic characterization, (3) suitability for computer processing, (4) lack of ambiguities, (5) inferential power, (6) explicit specification of rules for determination of language structure and content, (7) formal rule notation [25].



Medical Document: Collection of data representing medical informations with concern to a special patient. The types of data in a medical document are:

- identification data (patient identification, hospital identification, etc.),
- content data (lab results,(medical) language data, pictures, etc.).

Term: Word or phrase used for denotation of a semantic concept.

Descriptor: Term which can be used in an artificial language for denotation of a semantic concept. Descriptors can be ***simple*** descriptors or ***compound*** descriptors (denotation of compound semantic concepts).

Synonyms: Different terms denoting the same semantic concept.

Quasi-synonyms: Different terms denoting different semantic concepts which are not differentiated in the respective system.

Homonym: Term denoting context-dependent different semantic concepts.

Equivalence class: Set of all synonyms and quasi-synonyms.

Preferred term: Selected descriptor representing an equivalence class, i. e., representing a semantic concept. Sometimes the terms "descriptor" and "preferred term" are used as synonyms.

Nomenclature: List of terms for denotation of semantic concepts.

Terminology: List of terms related to semantic concepts.

Classification: Set of all preferred terms and of their (hierarchical) relations.

Thesaurus: List of descriptors and relations between descriptors:

- Used for ***terminologic control,*** i. e., translation from natural language utterances into artificial language utterances and vice versa.
- ***Controlled*** and ***dynamic vocabulary*** containing all terms, which are used to describe a specific domaine of discourse.

A thesaurus contains implicitly a classification, but is restricted to those terms and relations which are used. Additional weights of descriptors may be defined to identify the relevance of a descriptor for information retrieval.

Equivalence relation: Establishing of an equivalence class.

Hierarchical relation: Logical subordination, expecially

> Generic relation: The semantic concept denoted by the more specific term is fully contained in the semantic concept denoted by the more general term and differs from the more general semantic concept with concern to at least one characteristic (e. g., inflammation - acute inflammation).
>
> Partitive relation: The more specific term denotes a part of a whole, the latter being denoted by a more general term (e. g., kidney-urogenital system).

Associative relation: Each important relation which is neither an equivalence relation nor a hierarchical relation, e. g., marking of terms with the same generic general term (e. g., urogenital system - kidney, ureter, urinary bladder, ...) (see section 7.1).

## 3.2 TYPES OF MEDICAL UTTERANCES

The most important types of utterances with respect to medical language data processing are symptoms, signs, diagnoses, orders and descriptions.

The data collected in the usual medical history taking are subsumed here under "descriptions". In general, these descriptions lack in precision and standardization and they are on a low level of objectivity and abstraction. In pathology, for example, where the data seem to be more reproducable than in other medical areas, the descriptive part of a report serves several objectives:

- *Justification* of a diagnosis by verbalizing a picture and listing the criteria related to the diagnosis.
- *Modification* of a diagnosis when the diagnostic term available does not meet fully the findings.
- *Compensation* for a missing unique diagnosis.

Experience leads to the conclusion, that a diagnosis usually cannot be derived from descriptions alone, which is even true for that pathologist who formulated the description. In general, he prefers to look at the specimen. That means, a diagnosis contains more or other informations than the related description does. Therefore, and because descriptions are formulated in natural language in medicine, processing of pure descriptions will be excluded from the discussion for the present. But because there is no sharp distinction between medical language (e. g. diagnoses) and natural language in medicine, the problems may nevertheless occur and will have to be handled.



## 4 Morphology

Morphology is the theory of minimal meaningful units or morphemes and their arrangements in forming words. Morphemes are also called ***basic productive constituents***. Morphemes are strings of characters of minimal length, carrying grammatical and/or semantic information which is relevant in automatic language data processing. It is important to mention that morphemes are not only defined by a ***sequence*** of characters but also by their ***position*** in a sequence of morphemes forming a word. This can be demonstrated by the following examples (e. g., VERBIND<u>UNG</u> - <u>UN</u>GEBUNDEN, PERFORAT<u>ING</u> - <u>IN</u>GENIOUS). The order of morphemes forming a word is usually much more rigid than the order of words forming an utterance. This fact has been used to justify a contrast between ***morphology*** and ***syntax***. On the other hand, it can be argued that the inflectional morphemes in inflectional languages are markers not only of ***morphological*** relationships but also of ***morphosyntactic*** and ***morphosemantic*** relationships or values and indicate the boundaries of syntactic structures. For this reason, it does not seem to be useful to draw a sharp distinction between morphology and syntax.

For conveniance the following formal notation for case and number will be adopted: S denotes singular, P denotes plural and the digits 1, 2, 3 and 4 denote the different cases, e. g., S1 (nominativ singular), P14 (nominativ and accusative plural).

In morphosyntactical analysis a word w can be split into a ***stem form*** t and an ***inflectional morpheme*** s:

$$w = t * s.$$

In general, the stem form t carries the semantical information and the inflectional morpheme carries syntactical information:

$$w = \text{MAGENS} \longrightarrow t = \text{MAGEN}, \quad s = \text{S}.$$

The syntactical information of the morpheme s is not one-valued but is a function of both components of a word. This issue can be formulated by a bivariate function, where the same inflectional morpheme s may have different values:

$$f(\text{MAGENS}) = f(\text{MAGENS*S}) = \text{S2}$$
$$f(\text{TESTS}) = f(\text{TEST*S}) = \text{S2 P1234}$$

Because stems can be connected to different inflectional morphemes and the same inflectional morpheme can be connected to different stems, the number of different words is by far greater than the sum of the numbers of stems and inflectional morphemes. If the number of stems is $n_t$ and the number of inflectional morphemes is $n_s$, than the total number of components is $n = n_t + n_s$, whereas the product $n_t \cdot n_s$ is an upper limit for the total number of words. In natural languages the number of different words is much less then this upper limit, because most of the theoretical combinations of stems and inflectional morphemes are not used. But this number is still by far greater than the sum of the numbers of the components. For this reason, most projects in machine translation use morpheme dictionaries instead of dictionaries of full forms. Following this *segmentation* approach [27, 30, 41] we have several advantages:

- The dictionaries are relatively small, leading to substantial savings in *storage space* and *access time*.
- New words can be segmented, even if they have not been considered when the dictionary has been constructed. Therefore, necessary *dictionary update* is significantly reduced.
- About 2/3 of medical terms are based on Greek and Latin. Therefore, a dictionary of morphemes may serve as a basis for *multilingual dictionaries*.

These advantages are partly compensated by the need for the development of rules for the combination and disambiguation of morphemes. Especially a *discovery procedure* for the segmentation of words and a *synthesizer* for the construction of words or even phrases from morphemes have to be developed.

## 4.1 FORMAL REPRESENTATION OF RESULTS

There are several formal representations of the results of a morphosyntactical analysis in computable form, such as

- Tree representation (Fig. 1),
- Formation rules (Fig. 2),
- Matrix representation (Fig. 3),
- Decision table (Fig. 4).

## 4.2 MORPHOSYNTACTIC ALGORITHMS

So far, the notion of stems and inflectional morphemes has been introduced. German language and medical language make extensive use of compound word forms which may even paraphrase whole sentences. In medical language the reason for this behaviour is quite clear: utterances are shorter, simpler and therefore less ambiguous:

LARYNGOTRACHEOBRONCHITIS = Inflammation of the larynx, trachea and bronchus



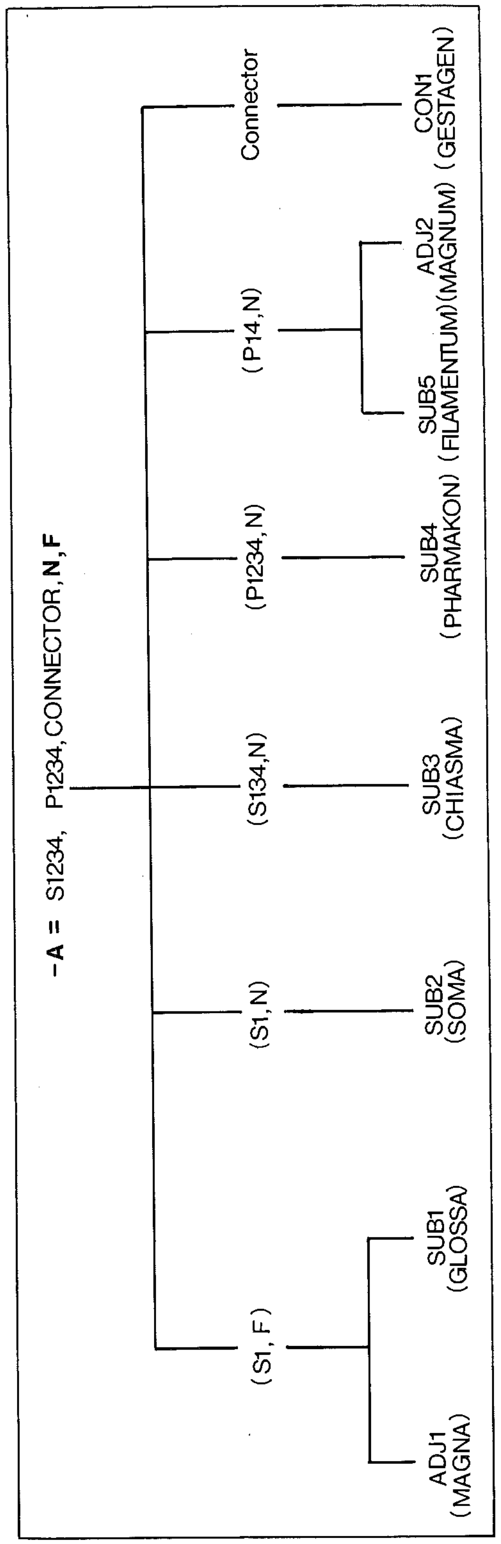



Fig. 1: Tree representation of the results of the morphosyntactical analysis of suffix -A in German medical language data. There are 2 adjectival classes (ADJ) and 5 noun classes (SUB). For each class an example is given. The genders are denoted as N for neutrum and F for feminine.

$$
\begin{array}{ll}
- A \cdot (SUB1 \cup ADJ1) & \subset (S1, F) \\
- A \cdot (SUB2) & \subset (S1, N) \\
- A \cdot (SUB3) & \subset (S134, N) \\
\ldots\ldots\ldots\ldots &
\end{array}
$$

Fig. 2: Formation rules for the results of the morphosyntactic analysis of suffix -A in German medical language data.

| | SUB1 | SUB2 | SUB3 | SUB4 | SUB5 | ADJ1 | ADJ2 | CON1 |
|---|---|---|---|---|---|---|---|---|
| S1 | F | N | N | | | F | | |
| S2 | | | | | | | | |
| S3 | | | N | | | | | |
| S4 | | | N | | | | | |
| P1 | | | | N | N | | N | |
| P2 | | | | N | | | | |
| P3 | | | | N | | | | |
| P4 | | | | N | N | | N | |
| Con-nector | | | | | | | | * |

Fig. 3: Matrix representation for the results of the morphosyntactic analysis of suffix -A in German medical language data.



| | SUB1 | SUB2 | SUB3 | SUB4 | SUB5 | ADJ1 | ADJ2 | CON1 |
|---|---|---|---|---|---|---|---|---|
| -A | (S1, F) | (S1, N) | (S134, N) | (P1234, N) | (P14, N) | (S1, F) | (P14, N) | * |
| -AE | . . . | | | | | | | |
| . | | | | | | | | |
| . | | | | | | | | |
| . | | | | | | | | |

Fig. 4: Decision table for the results of the morphosyntactic analysis of suffix -A in German medical language data.

The paraphrase uses 7 words, a comma and 41 characters other than blank, whereas the compound word form uses only 24 characters. The compound word form carries additional information, because the sequence of the site markers follows a logical path. It may therefore be more efficient to segment stems into morphemes.

Two models may be presented here. In one model terminal morphemes are identified, in the other one there is a full segmentation of words.

### 4.2.1 Identification of Terminal Morphemes

Suffixes are analyzed from right to left and are compared with a tree-structured set of "productive" suffixes [30]. The longest entry found in the tree, matching the analyzed suffix, is taken. The length of suffixes is restricted to 4 characters. Productive suffixes are connected to a set of rules, identifying parts of speech, as well as

- adjective to noun transformations,
- noun to noun transformations,
- noun plural to noun singular transformations.

Parts of a rule are <type> and <remainder>. <type> identifies a suffix as adjective (A) or noun (N). The remainder consists of permissible morphological transformations (Fig. 5). If necessary, the transformations are applied and are verified using a reference dictionary. A permissible transformation consists of a target parts of speech symbol, an integer, denoting the number of characters which have to be deleted and a

| | |
|---|---|
| L | |
| AL | A N2A N2E N2UM N2US |
| NAL | |
| INAL | A N4EN |
| EAL | A N3US N3ES |
| GEAL | A N4X |
| CEAL | A N4X |
| IAL | A N3US |
| CAL | A N4IX N4EX N4ERY |
| ICAL | A N4IX N4EX N4ERY |
| RAL | |
| ORAL | A N4UR |

**Fig. 5:** **L-tree of productive suffixes [30].**

string, which has to be appended to the right of the remaining string. The model will be discussed in more detail in section 8.4.1.

### 4.2.2 Segmentation Approach

The second model uses a segment dictionary [41] and will be discussed in more detail because it seems to be more efficient for medical language.

Ultimate goal is the segmentation of every word used in medical language into pieces (*segments*) carrying syntactical and/or semantical information which cannot be broken down into smaller pieces. The linguistic entity "morpheme" often imposes problems, which may be avoided. Therefore, an approach is taken, which is slightly different from the usual linguistic approach.

The principles underlying the model are:



- There are different types of segments, listed in dictionaries; the dictionaries are linked by a set of rules.
- There is an explicit structure of words as a repetitive sequence of elements, where each element consists of segments of different types in fixed order.
- Segmentation of words is done from left to right and follows a principle of longest valid match.

#### 4.2.2.1 Words and Segments

For further description there have to be definitions of different types of *segments*.

A *root* is a string of characters carrying semantical information which cannot be decomposed into meaningful parts by splitting the root. Note, that not every string, which carries semantical information has to be a root. In practice, the definition is done by enumeration.

A *secondary string* is every string which is not part of a root and which is used for building a word. The empty string is a secondary string. In general, a secondary string consists of two parts:

(1) *derivational* suffix,
(2) *terminal* suffix.

The terminal suffix can be an *inflectional* suffix or it can be a *connector* (Fig. 6).

$$\boxed{w = R_1 s_1 t_1 R_2 s_2 t_2 \; . \; . \; . \; R_n s_n t_n \quad (n \geq 1)}$$

**Fig. 6:** **Canonical form of a word w in terms of roots (R), derivational suffixes (s) and terminal suffixes (t).**

For simplicity, I will not differentiate between the two parts of a secondary string in the following discussion.

With some language specific exceptions there is:

- The derivational suffix determines the connector(s) and the inflectional suffixes.

- In many noun-adjective transformations only the exchange of the secondary string is necessary.
- Many roots are derived from Latin or Greek roots, sometimes used in their original form, sometimes adapted to the English or German language. In the latter case, mostly the replacement of language specific parts of the secondary string is necessary for translation from English to German and vice versa.

Sometimes suffixes carry semantic informations. This has been analyzed for the suffixes *-osis* and *-iasis* in English and French [27]. Other examples are: *-itis* (inflammation), *-oma* (tumors), *-ase* (enzymes), *-one* (ketones), *-ite, -ide* (salts). In these cases the decision is arbitrary whether the segment belongs to the set of roots, or to the set of secondary strings carrying semantic information. So far, these segments are used as secondary strings due to a more or less technical reason: These segments are always preceded by roots. If they would be defined as roots, than the number of roots, which can be connected by the empty string would increase and the degree of ambiguity would be raised.

A ***morphosyntactical class*** number n contains a set (family) $S_n$ of secondary strings

$$S_n = \{s_{n1}, s_{n2}, \ldots\}, \quad 0 \leq n \leq 255,$$

where each member $s_{ni}$ is associated with a specific syntactical and/or semantical information.

For a root R associated with class number n all the constructions

$$R * S_n = \{Rs_{n1}, Rs_{n2}, \ldots\}$$

are ***valid***, that is, the sequence $Rs_{ni}$ is correct in a word. Each construction $Rs_{ni}$ carries the syntactical information associated with $s_{ni}$.

If a secondary string $s_j$ is associated with class number n, than $s_j$ is a member of family $S_n$.

There are about 250 classes with morphosyntactical (Fig. 7) and semantical information. Each root is assigned to at least one class. The maximum number of classes a root belongs to is 14.

Morphosyntactical and semantical values are assigned to constructions and are not related to secondary strings alone. Thus there is no problem to distinguish between suffix -om marking a tumor and suffix -om in Coelom (Fig. 16).



| | Case | Number | Gender | Degree of comparison | Person | Tense | Voice |
|---|---|---|---|---|---|---|---|
| Nominals | * | * | * | | | | |
| Pronominals (excl. personal pronouns) | * | * | | | | | |
| Adjectivals Participles | * | * | * | * | | | |
| Verbals | * | * | | | * | * | * |
| Personal pronouns | * | * | | | * | | |

**Fig. 7:** **Syntactical informations for different parts of speech.**

The purpose is the same information PRATT [30] gets by the identification of terminal morphemes. The advantage is the handling of smaller pieces of information and the knowledge of the root and secondary string usable with the respective root, i. e., there is a direct relation between the various possible forms derived from a root, thus reducing the number of noise words generated significantly and establishing a direct relation between the original Latin or Greek form and the equivalent English or German form in about 2/3 of all cases. This is a basis for the translation of diagnostic statements. The disadvantage is the need for a larger dictionary.

Detailed class descriptions go beyond the purpose of this paper, but a few examples may show the principles (Fig. 8 - 12).

A construction $d_i d_{i+1}$ is a ***valid pair***, if $d_i$ is a root and $d_{i+1}$ is a secondary string and $d_i$ and $d_{i+1}$ are associated with the same class number (there may be more than one).

A sequence $d_1 d_2 \ldots d_j$ is a ***sequence of valid pairs***, if each pair $d_i d_{i+1}$ (i being an odd number) is a valid pair and if - for an odd number j - $d_j$ is a root.

The ***canonical form*** of a word w is a sequence of valid pairs (Fig. 6)

$$w = R_1 s_1 R_2 s_2 \ldots R_n s_n .$$

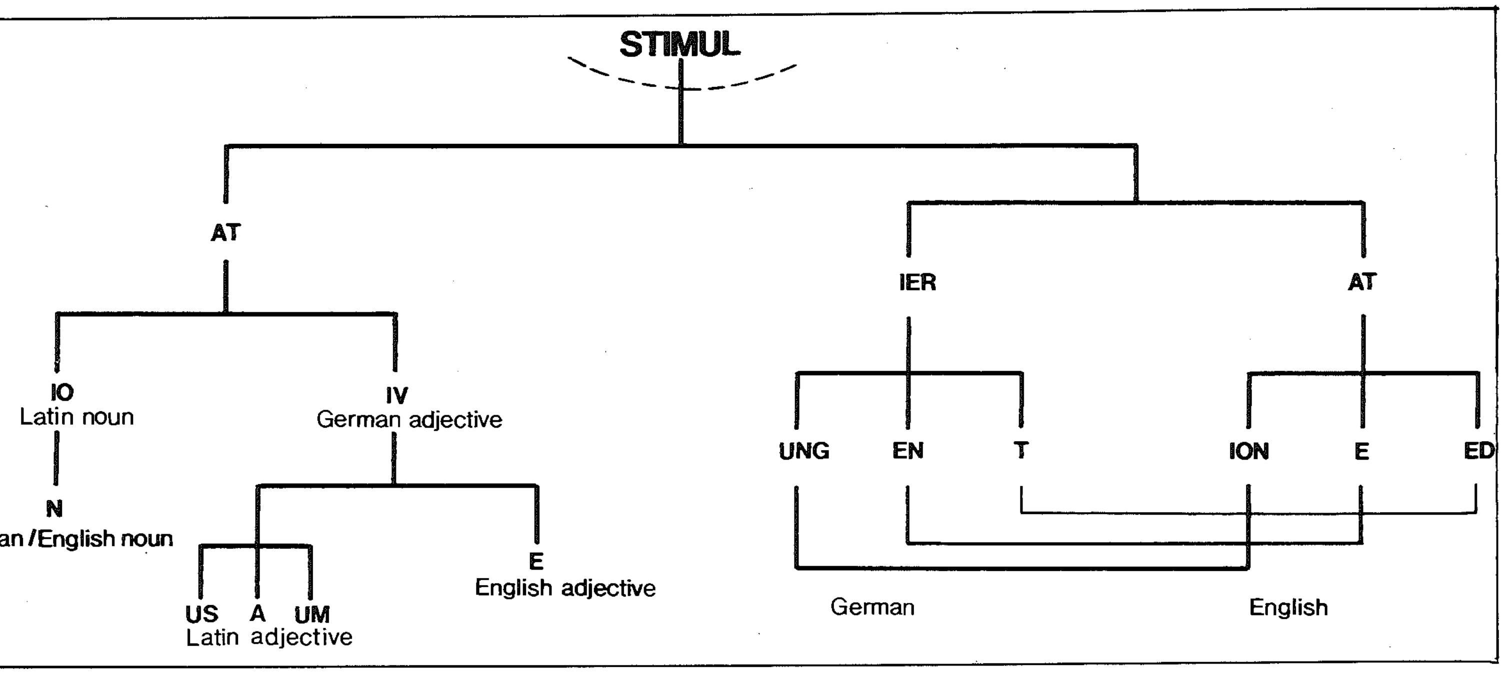



Fig. 9: Suffix families $S_{31}$ and $S_{86}$[41].



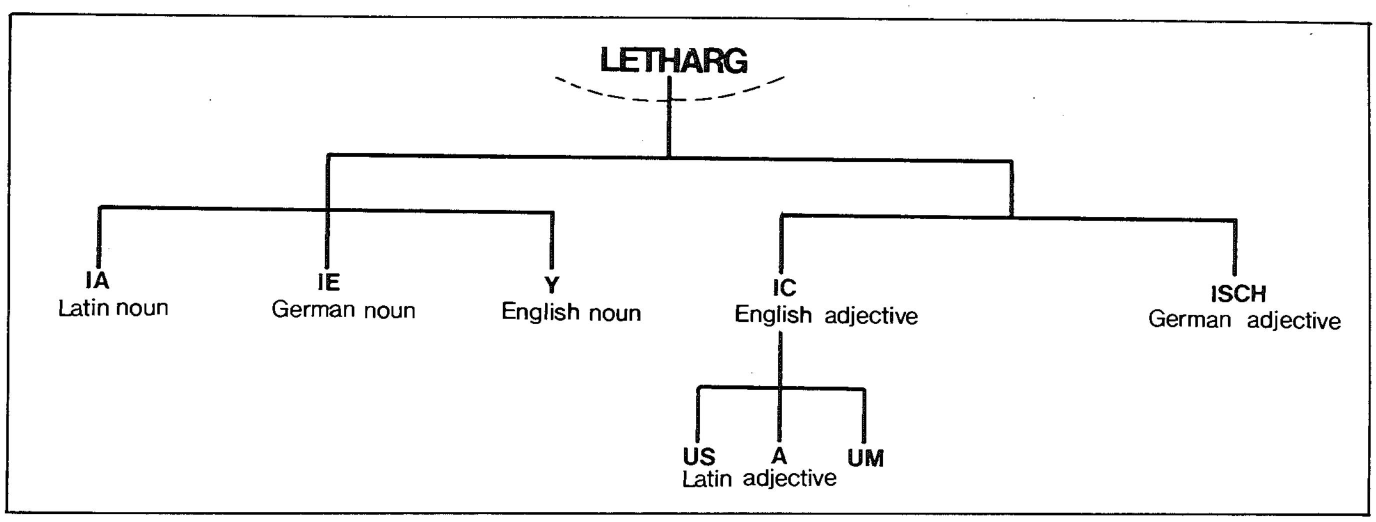



Fig. 10: Suffix families $S_{33}$ and $S_{66}$[41].

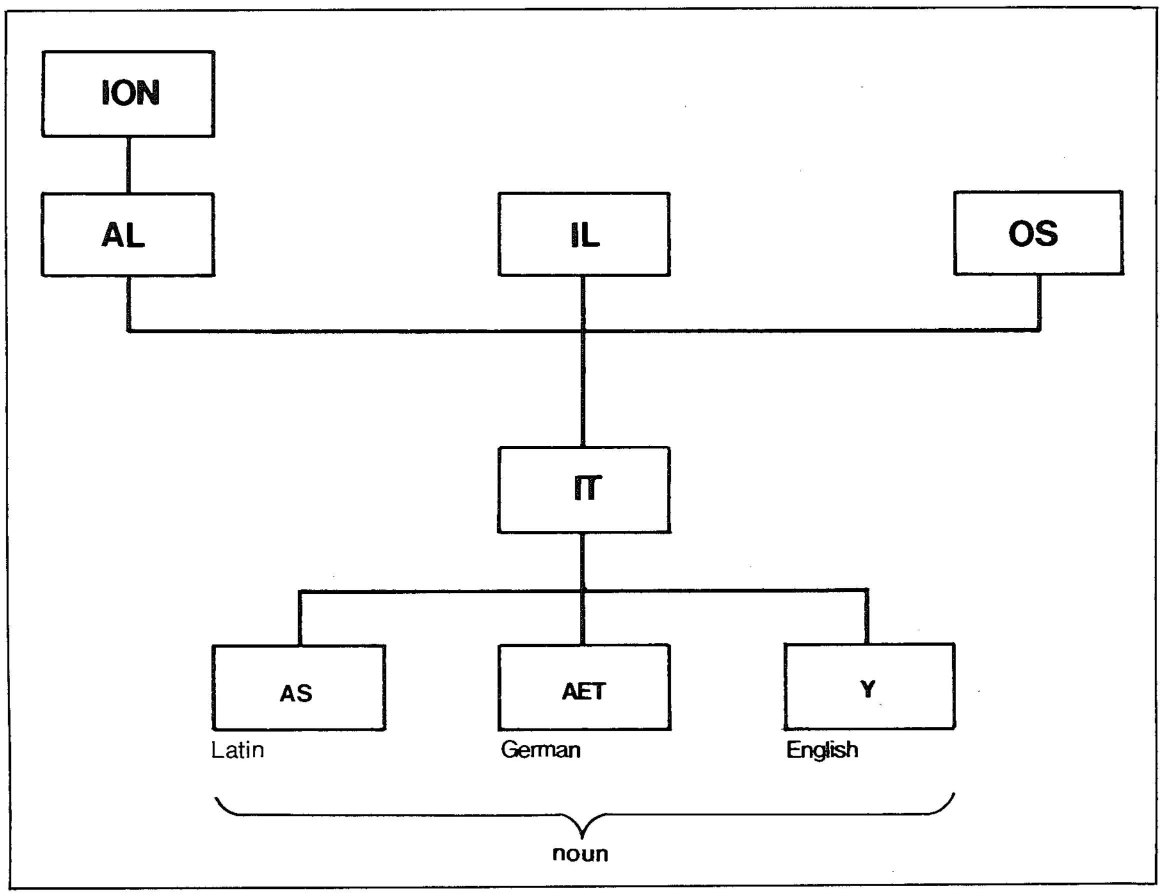



Fig. 11: Part of a suffix tree. Each box belongs to another family [41].

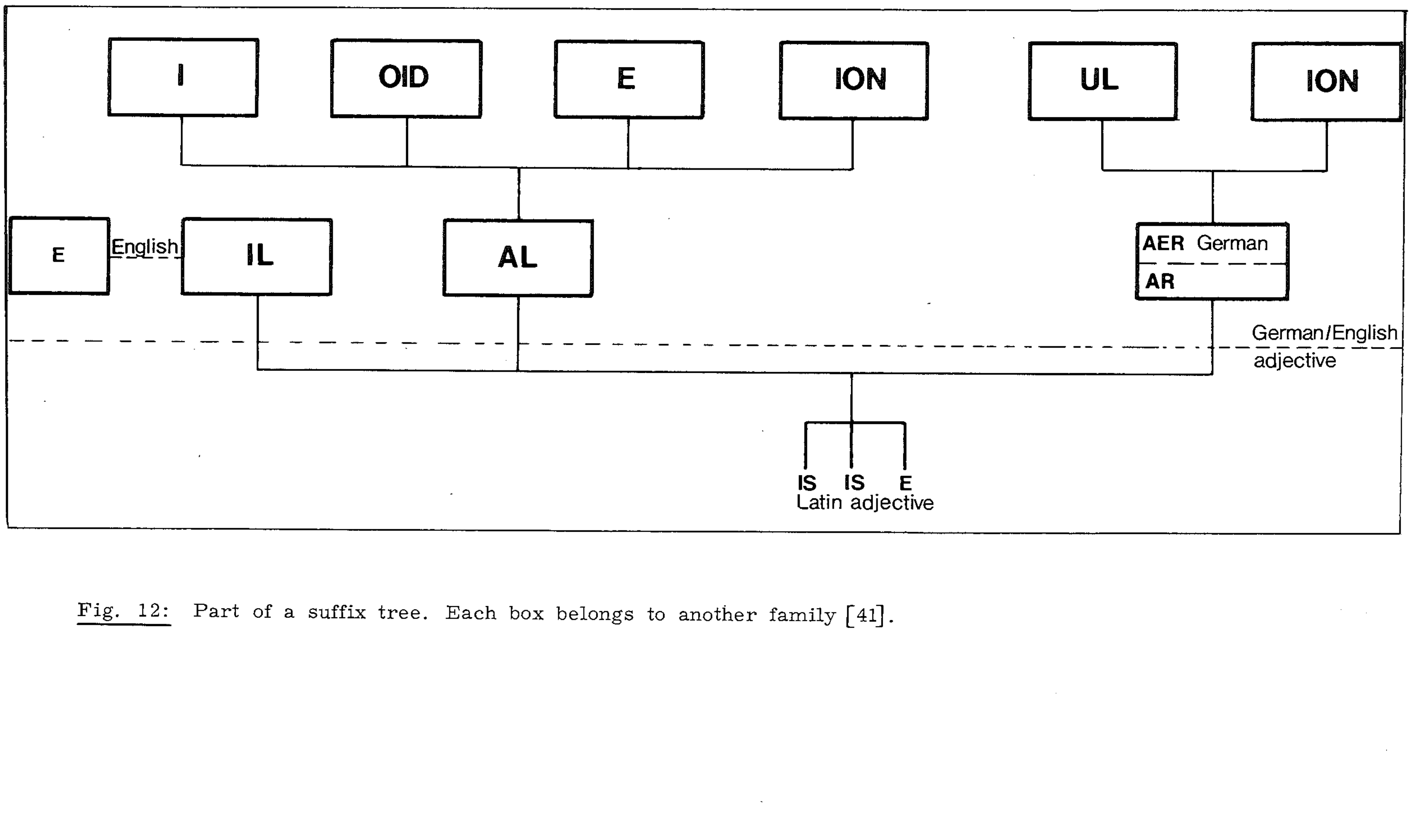



Fig. 12: Part of a suffix tree. Each box belongs to another family [41].

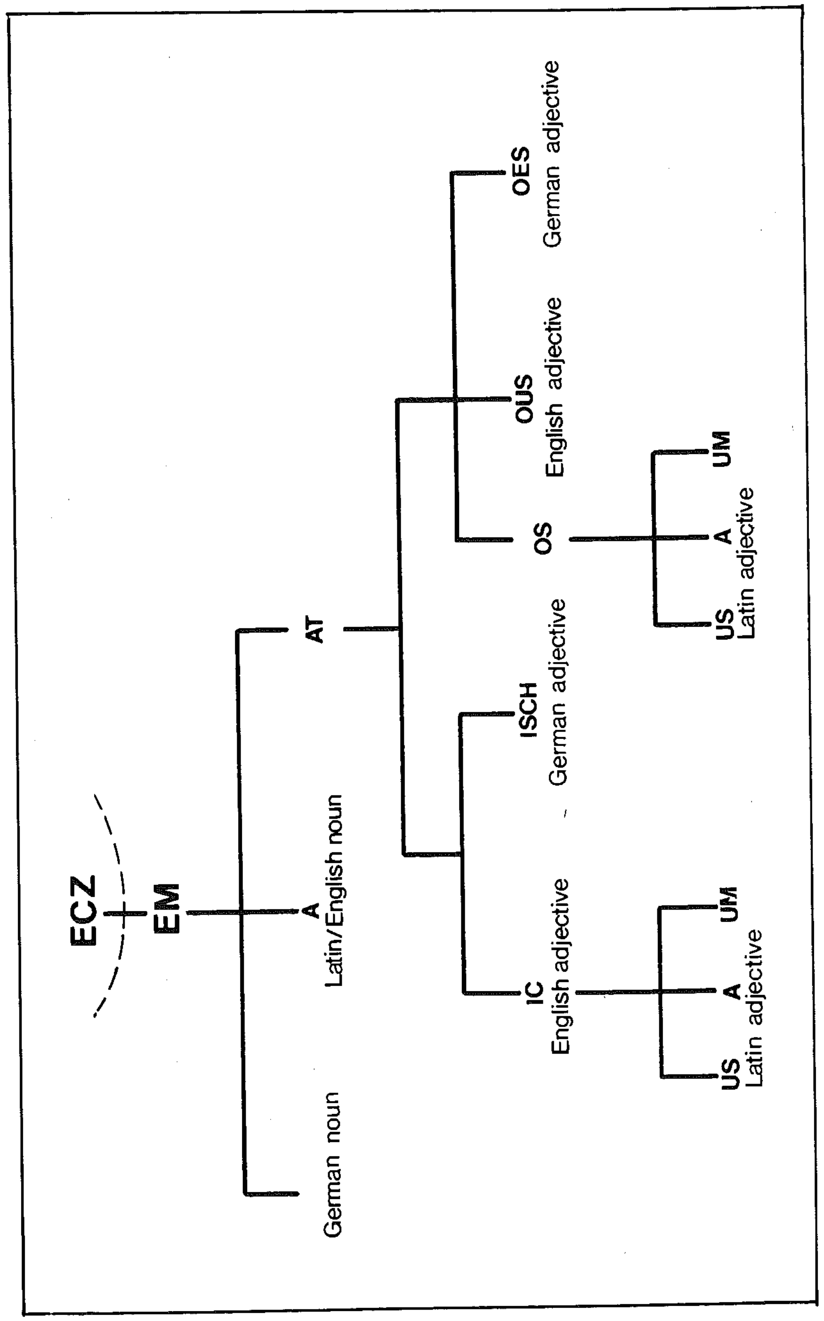



Fig. 8: Suffix family $S_{80}$[41].



There are a few additional rules, e. g., segments which must not be terminal segments or segments which have to be terminal segments.

Preliminary analyses showed, that the similarities proven for English and French in special examples hold also for German in the majority of cases. Therefore a German segment dictionary has been constructed semi-automatically by use of the segmentation algorithm (see section 4.2.2.3) and starting from a word-oriented German thesaurus [32] containing about 28,000 different words from autopsy and biopsy reports. To this about 12,000 different words from a German SNOP translation [42] have been added. Then the resulting German segment dictionary has been used for the segmentation of about 8,000 different words from the English SNOP version [38], not considering chemicals and drugs. Without any addition to the dictionary about 70 % of the English words could be segmented automatically. This number increased to more than 80 % after addition of about 20 specific English suffixes. The rest consisting of specific English roots and secondary strings has been added manually (Table 1).

| Number of | Roots | Secondary strings | | Segments | % |
|---|---|---|---|---|---|
| | | Derivational suffixes | Terminal suffixes | | |
| German | 1583 | 122 | 7 | 1712 | 21 |
| English | 814 | 111 | 1 | 926 | 12 |
| German and English | 5139 | 179 | 52 | 5370 | 67 |
| Total | 7536 | 412 | 60 | 8008 | 100 |

Table 1: **Statistics for a segment dictionary for medical language data processing.**

4.2.2.2 Organization of the Segment Dictionary

Because of the extensive intersection, English and German segment dictionaries are merged and each segment is assigned a ***language marker*** L, which identifies, whether it belongs to the German, to the English or to both versions of the dictionaries.

The general form of an element of the dictionary is

$$\begin{Bmatrix} R \\ S \end{Bmatrix} L\ i_1 \ldots i_j \ldots i_k, \quad 0 \leq i_j \leq 255,\ j = 1,2,\ldots,k, \quad k \leq 14 ,$$

where R and s are strings of characters (root, resp. secondary string), L is the language marker and i is a class number. Every construction of $R * S_n$ is possible ($n = i_1, \ldots, i_k$), that is, every construction Rs is valid, where s belongs to the union of all sets $S_n$ (Fig. 13).

### 4.2.2.3 The Segmentation Algorithm

The segmentation algorithm is searching ***alternatively*** for roots and secondary strings in a word. The principles of the algorithm are: There are two dictionaries, a dictionary $D_r$ of roots and a dictionary $D_s$ of secondary strings. The segments are assigned class numbers. Both dictionaries are in lexicographical order. Thus the relation $d_1 < d_2$ means, $d_1$ stands before $d_2$ in lexicographical order. The empty string belongs to $D_s$. Segments are searched from left to right. Given a sequence

$$w = d_1 d_2 \ldots d_i X ,$$

$d_1 \ldots d_i$ being a sequence of valid pairs and X the remainder of word w, the algorithm is searching for the ***longest*** segment $d_{i+1}$, such that

$$w = d_1 d_2 \ldots d_i d_{i+1} Y ,$$

and $d_1 \ldots d_{i+1}$ is a sequence of valid pairs (***principle of longest valid match***). If such a segment is found the algorithm proceeds for Y. Otherwise, a mistake is assumed in the preceding sequence. First $d_i$ is assumed to be wrong, the correct sequence then being

$$w = d_1 \ldots d_{i-1} t_i Z ,$$

where $t_i < d_i$ and $d_1 \ldots d_{i-1} t_i$ being a sequence of valid pairs. If such a segment $t_i$ is found, the algorithm proceeds to the right. Otherwise $d_{i-1}$ is assumed to be wrong, etc..

The algorithm terminates when either the whole word has been segmented, i. e., $w = d_1 \ldots d_n$ is in canonical form, or there is no root as first root.

The basic assumptions underlying this model are:

(1) The semantic content of a word is the logical sum of the semantic contents of its roots. This is the most simplistic assumption, which can be made to reduce the size of any semantic structured reference dictionary. A much more complicated set of rules has to be developed for the prediction of the value of compound words, where this simplistic model is insufficient (Examples [26] may demonstrate some aspects (Fig. 14)). In many cases, the semantic category of the compound descriptor is identical to the semantic category of the last segment.



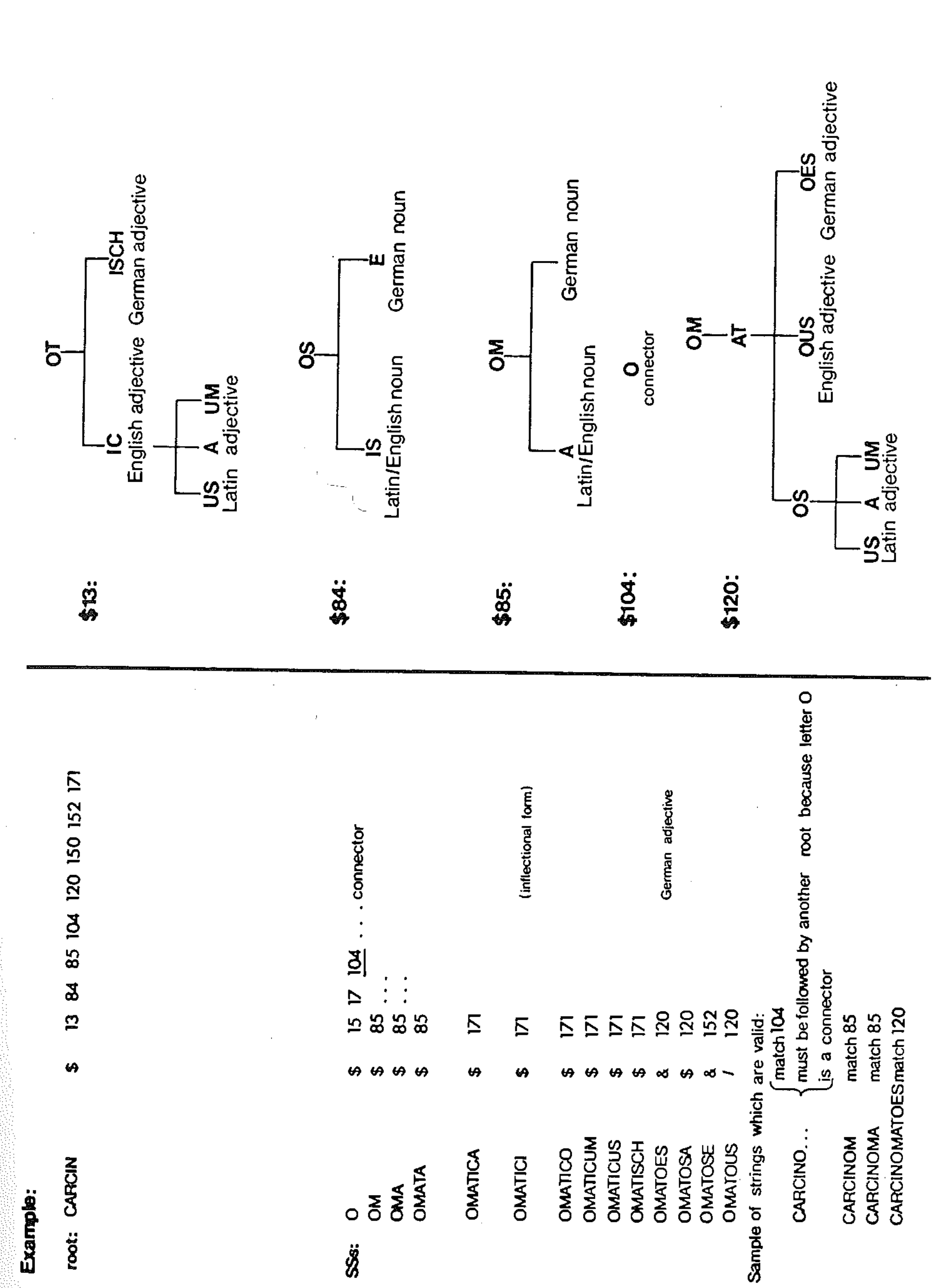



Fig. 13: Part of an entry into the root dictionary and some related entries into the dictionary of secondary strings ($SS_s$) on the left side. Some suffix families related to the root on the right side [41].

| Semantic Category of Constituents | Graphemic Segmentation of Word Form | Semantic Category of Word Form |
|---|---|---|
| T T | MYO, CARDIUM | T |
| T M | BRONCH, ITIS | M |
| T E | MENINGO, COCCUS | E |
| T F | VASO, SPASM | F |
| M T | VARICO, BLEPHARON | M |
| E F | AMEB, IASIS | F |
| E E | BACTERIO, TOXIN | E |
| F F | ALGO, SPASM | F |
| T T M | GLOMERULO, NEPHR, ITIS | M |
| F T F | HYPER, TRICH, OSIS | M |
| T E F | THYRO, TOXIC, OSIS | F |
| F T M | MENO, METRO, RRHAGIA | F |
| T T T T M M | PERINEO, COLPO, RECTO, MY, OM, ECTOMY | M |

Fig. 14: Semantic patterns of compound words and resulting semantic values [26]. The semantic categories due to SNOP [38] are: T = topography, M = morphology, E = etiology, F = function.

| | | |
|---|---|---|
| SCHLÜSSELBEIN | | |
| LANDKARTENZUNGE | - | GEOGRAPHIC TONGUE |
| SANDUHRMAGEN | - | HOURGLAS STOMACH |
| KEIMEPITHEL | | |
| NARBENHERNIE | | |
| SCHUSTERBRUST | - | COBBLER'S CHEST |
| CYSTENLUNGE | - | HONEYCOMB LUNG |
| NIERENBECKEN | - | RENAL PELVIS |
| BECKENNIERE | - | PELVIC KIDNEY |

Fig. 15: Examples where decomposition of compound descriptors lead to changes in the semantic concepts.



(2) The parts of speech can be derived from the bivariate function $f(R_n, d_n)$ using last root $R_n$ and ***derivational*** suffix $s_n$.

(3) Case, gender, etc. can be derived from the three-variate function $g(R_n, s_n, t_n)$ using last root $R_n$, derivational suffix $s_n$ and terminal suffix $t_n$. There is a substantial number of derivational suffixes, where this information can be derived from a bivariate function using only the secondary string.

There are several exceptions to these basic rules, which have to be handled separately. Types of exceptions to rule (1) are:

- Words with a semantic meaning differing from the sum of the semantic meanings of the constituents (Fig. 15).
- Ambiguous semantic meaning of components (MYEL - referring to "bone marrow" or to "spinal cord", SCLER - referring to SCLERA or SCLEROSIS).
- Secondary strings carrying a specific semantic meaning, like -OM, -OMA (tumor) (Fig. 16), -IASIS (diseases), -ITIS (inflammation) except capitis, linitis plastica, rhachitis.

| | |
|---|---|
| SYNDROM | FIBROM |
| SKOTOM | SPIDER ANGIOM |
| COLOBOM | OSTEOM |
| HAEMATOM | TRACHOM |
| GRANULOM | GLAUKOM |
| TUBERCULOM | MYCETOM |
| STAPHYLOM | RHINOSKLEROM |
| LEUCOM | PARAFFINOM |
| NEUROM | CONDYLOM |
| ATHEROM | KERATOACANTHOM |
| XANTHOM | FIBROMA PENDULUM |
| PSAMMOM | ANGIOMA SERPIGNOSUM |
| CHOLESTEATOM | OSTEOFIBROM |
| PAPILLOM | COELOM |

Fig. 16: Exceptions to the $\left\{\begin{matrix}\text{-OM}\\ \text{-OMA}\end{matrix}\right\}$ rule with respect to SNOP [38].

The semantic meaning of suffixes often is not connected to the suffix alone but is a bivariate class function. For example, the suffixes -IA (Latin), -IE (German) and -Y (English) constitute two classes with the same morphosyntactical but with different semantical information:

$$f(\text{FAMIL*}\begin{Bmatrix}\text{IA}\\\text{IE}\\\text{Y}\end{Bmatrix}) = \text{noun, singular}$$

$$f(\text{ANODONT*}\begin{Bmatrix}\text{IA}\\\text{IE}\\\text{Y}\end{Bmatrix}) = \text{noun, singular, disease}$$

#### 4.2.2.4 Synthesis rules

Insofar, we have used the classes and analysis rules for the segmentation of compound word forms. The same rules can, however, be used for synthesis and for transformation. This has been made possible by two different features:

- combinatorial classes consisting of subclasses for noun form, adjectival form and so on (Fig. 8),
- membership of a root in several classes, e. g.

$$\text{ENTZUEND} < \{\text{UNG}\} \cup \{\text{-EN}\} \cup \{\text{-ET}\} \cup \{\text{-LICH}\}.$$

The following transformations are possible:

- Noun singular $\longleftrightarrow$ noun plural,
- noun $\longleftrightarrow$ adjectival,
- noun $\longleftrightarrow$ verbal,
- German noun $\longleftrightarrow$ English noun $\longleftrightarrow$ Latin noun, if possible using class functions.
- German adjectival $\longleftrightarrow$ English adjectival $\longleftrightarrow$ Latin adjectival, if possible using class functions.

#### 4.2.2.5 Eponyms

Eponyms are used extensively in medicine and have to be treated specially. Eponyms mostly are

- names of people (e. g., HODGKIN'S DISEASE, MORBUS LITTLE),



- names of geographical regions (e. g., CHRISTMAS DISEASE, COXSACKIE VIRUS),
- names of professions (e. g., WOOLSORTER'S DISEASE, COBBLER'S CHEST).

The eponyms have to be listed in their full forms in the root dictionary and they make up a good portion of it. Additionally some of these eponyms are inflected like Latin roots (e. g., FALLOPIAN TUBE, TUBA EUSTACHII, ANGINA LUDOVICI).

#### 4.2.2.6 Problems

Some of the problems arising from the fact that there are only a few principles underlying the model have already been mentioned. Therefore, I will concentrate on some special failures, which are of practical interest.

Ambiguities have to be handled, e. g., the word ECTOPIA can be segmented formally into the segments

$$\underbrace{\text{EC}}_{R_1}\ \underbrace{\emptyset}_{s_1}\ \underbrace{\text{TOP}}_{R_2}\ \underbrace{\text{IA}}_{s_2} \quad\text{and}\quad \underbrace{\text{ECT}}_{R_1}\ \underbrace{\text{O}}_{s_1}\ \underbrace{\text{P}}_{R_2}\ \underbrace{\text{IA}}_{s_2}\ .$$

The first sequence is the correct one, but the second sequence is the canonical form due to the principle of longest valid match (the root dictionary contains the longer prefix ECT, which may be connected by letter O, e. g. ECTODERM). This problem arises from the unsymmetrical segmentation algorithm and it is not detected automatically. It may occur whenever roots are combined with some "reactions" in the combination area, e. g.,

(1) elision of vowels (MYOPIA instead of MYOOPIA),
(2) "smoothing" of combinations (ECTOPIA instead of ECTOTOPIA),
(3) dropping of one consonant in case the preceding root ends with the same (double-)consonant the following root starts with (FLUSS-STAUUNG $\longrightarrow$ FLUSSTAUUNG),
(4) gemination (TAR $\longrightarrow$ TARRING).

Errors like the example given occur in a surprisingly small number of less than 3 % in a large sample of more than 50,000 different words. In most of these cases the ambiguities are caused by short prefixes, like A-, E-, and so on. There are different types of failures, which can be arranged into classes.

Practically all ambiguities could be detected, if the same principle of longest match would be applied not only in a left to right scan, but also in a right to left scan. But because it is intended to develop a segmentation algorithm, which can be used fully automatically, there is no advantage in the knowledge of more than one possible formal segmentation, as long as there are no explicit rules for the decision, which of the different versions should be taken.

The ambiguities can be handled by special rules or by addition of pseudo-roots to the dictionary. Where the number of cases is small the latter has been done.

## 5 Syntax

Syntax deals with the order or structure of word strings with respect to the parts of speech. I will not go into the discussion about the usefulness of special types of grammars, which is much more up to a linguist. It has often been argued, that linguistics has not yet delivered powerful tools for practical retrieval systems. I do not know, whether this is because of the insufficiency of formal grammars for natural languages or because of the insufficiency of people working with practical retrieval systems with a very limited knowledge of linguistic theories.

Therefore I will mention some types of formal grammars very briefly to introduce just the notions. ***Formal grammars*** intend to develop a formal model of a language and are therefore most relevant to automatic language data processing. The definition of language is by generation, using ***rewrite rules:*** "Rewrite a string X as a string Y". Such a grammar is adequate , if it generates just all sentences of the language. If more sentences are produced, than rules have to be more restrictive. Formal grammars are totally restricted to formal relationships and disregard semantic content.

A ***phrase-structure*** grammar consists of a finite ***vocabulary,*** V, and a finite set, $\{P\}$, of ***rewrite rules.*** The vocabulary, V, consists of two disjoint parts, a ***terminal*** vocabulary, $V_T$, and a ***non-terminal*** vocabulary, $V_N$. The terminal vocabulary is a lexicon of the language, the non-terminal vocabulary contains symbols, such as S = sentence, NP = noun phrase, VP = verb phrase, etc..

Each rewrite rule has the form

$$\alpha A \beta \longrightarrow \alpha \gamma \beta ,$$

where A is a member of $V_N$ and $\alpha, \beta$ and $\gamma$ are members of V. $\alpha$ and $\beta$ may be null, A may be not. This general form of a phrase-structure grammar is called ***context-sensitive.*** If the rewrite rules have the special form

$$A \longrightarrow \gamma$$

then the grammar is called ***context-free.*** A special context-free grammar is a ***finite-state*** grammar, with rewrite rules

$$A \longrightarrow cB$$
$$A \longrightarrow c$$



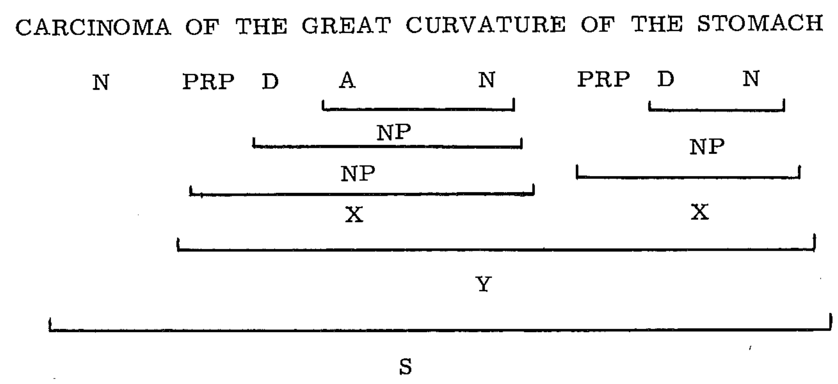
where A and B are members of $V_N$ and c is a member of $V_T$.

A finite state grammar is the most restricted phrase-structure grammar, which can generate an infinite number of sentences with a finite set of rewrite rules.

Another more restricted grammar may have only rewrite rules of the second type

$$A \longrightarrow c.$$

A grammar like this one is essentially a finite list of phrases, which constitute the language (Fig. 17).




**Rewrite rules**

S → N + Y
Y → X + X
X → PRP + NP
NP → D + NP
NP → N
N → stomach, curvature, carcinoma
D → the
PRP → of
A → great

**Fig. 17:** Example of the syntactical analysis of a medical utterance.

Another type of grammars is called *transformational* grammar [4]. Two structures are assigned to each sentence: A *surface* structure and a *deep* structure. The latter one determines the semantic interpretation.

## 6 Semantics

The different types of medical utterances have to be seen in the light of the semantic structure of informations in medicine. This semantic structure is always preliminary due to the limited knowledge about the universe. Formal representation of informations requires a model of the universe, which contains a finite number of elements:

- *Characteristics* and their respective *values* (semantic concepts),
- *relations* between different characteristics and values,
- a *terminology* containing the terms denoting the characteristics and their respective values,
- a *classification* to make explicit the relations between characteristics and values.

That is, modelling the universe leads directly to the problems of classification and terminology. Because models are used for a special purpose and depend on diagnostic and therapeutic possibilities, classification and terminology are a valuable resource but they cannot be the goal. The classification used depends on the goals and there is a variety of such classifications in medicine, a fact, which often leads to serious problems in communication instead of solving these problems.

There are efforts to *standardize* terms with respect to their meaning, which is an important requirement for every kind of language processing. For example, the terms "breathlessness" and "dyspnea" are often used synonymously to describe shortness of breath of normal individuals as well as of individuals suffering from pulmonary disease. Agreements in the medical community have to be made, which are the basis for unambiguous usage of terms.

The *Current Medical Information Terminology* (CMIT) [11] contains about 3,500 diseases and their definitions in terms of etiology, symptomatology, and X-ray and laboratory findings. In another system developed in Germany [20] data about rare or newly described syndromes are collected. There are efforts for the standardization of syndrome and symptom nomenclature. Further efforts are mainly in the area of tumor nomenclature [35, 36, 37].

Diagnostic terms are a good example to demonstrate the difficulties. In a common sense the diagnosis serves for different purposes, e. g., unambiguous and short denotation of a disease, explanation of disturbances, classification due to a given scheme. Be-



cause most of the diseases are complex in nature with respect to etiology, pathogenesis, combination of signs and symptoms and with respect to their time dependance and individual reaction, simplification has to be done. The classification therefore has to be - at least in principle - *polyhierarchical*, it has to be an *open system* to allow for additions, deletions and update. The necessary degree of *completeness* depends on:

- data,
- structure of search questions,
- handling by computer or by man.

The semantic categories, which are used to categorize medical informations can be explained, at least in part, by the history of medical systematization. I will not go into the interesting, but sometimes philosophical discussions about the value of a diagnosis. This has been done and this discussion is carried out by experts in the field [14]. From a pragmatic point of view, the semantic categories should meet - at least approximately - the following objectives:

- *Decomposition* of information, i. e., the elements should be disjoint and complete in the sense of set theory.
- Orientation towards the *diagnostic* and *therapeutic possibilities*.
- Differentiation between different *prognostic entities*.
- Reflection of known important *relations* and *pathogenetic processes*.

Some of the categories used in different systems are:

- Topography (location of a disease),
- nosology (disease entities),
- morphology (what can be seen, for example using tissue specimens and a microscope),
- etiologic agents (causes of alterations),
- functional disturbances,
- pathogenetic processes,
- histochemical alterations,
- medical procedures (diagnostic procedures, therapeutic procedures).

This list is far from being complete and it contains categories, which are not mutually exclusive. For example, the category "nosology" is based on the assumption that diseases are well definable states, which can be arranged in some type of a natural system in analogy to the natural system of elements [14]. In the light of the discussion whether this is right or wrong, terms like "disease" or "syndrome" already reveal different points of view. Without sticking to the assumption of the existence of a natural system of diseases, one can say that morphology, etiology, function etc. are semantic subdivisions of the category "nosology".

Diseases usually have projections into several semantic categories, e. g., *acute appendicitis* as a patho-anatomical diagnosis is made up from acute inflammation (morphology) + appendix (topography), whereas the same term as a clinical diagnosis is made up from abdomen (topography) + pain, nausea, emesis, fever (function).

In a consequent classification - if ever one such classification exists - one has to consider

- *multiple causes* of a disease or state,
- projections into *several semantic categories,* depending on the point of view and on the available facilities and knowledge,
- *inconstancy* of the projections depending on time and patient characteristics.

Taking into account these considerations, available systems are rather poor. Nevertheless there are developments, which are an important step forward.

The semantic categories, which have to be used, depend on the special demands in information retrieval and on the available resources with respect to dictionaries, classifications, thesauri, data and software.

It is agreed that more sophisticated language processing requires semantically ordered *lexicons.* These lexicons list the morphemes, words or phrases, which have some value of meaning in the respective domain of discourse. This value is identified by *semantic markers,* which reveal the meaning-defined use of these elements in terms of syntax, denotation, connotation, etc. [29]. In addition, thesaurus functions should be preserved, like:

- Entry by letter strings to yield concepts, and
- entry by concepts (semantic markers) to yield lists of words and phrases related to the concepts.

There are three general approaches to the construction of a semantic lexicon.

The *scaling* method uses judgements along semantic dimensions from language users about the meaning.

The *statistical derivation* method examines the contextual distribution of two words



or phrases to determine their similarity of meaning.

The *componential analysis* decomposes the sense of each word or phrase into a number of components. This is done by assigning to each word-sense in the dictionary:

(1) A set of semantic markers, which are *overall* semantic features, occuring in more than one word sense.

(2) A *specific* marker, which represents the semantic features unique to the given word sense.

(3) *Combinatorial rules,* that are the conditions under which the given word-sense may be combined with other word senses.

Additionally there are *projection rules* describing the methods of combining the readings of a pair of words or phrases to obtain a reading for the longer string.

The above listed requirements are best met by *thesauri,* possibly completed by a set of special dictionaries (*microglossaries*). It has already been mentioned, that a thesaurus is a list of *descriptors* and *relations* between descriptors, containing all the terms, which are necessary to formulate the semantic concepts in a domain of discourse. The descriptors can be divided into a set of simple descriptors and a set of compound descriptors. A *simple descriptor* denotes a semantic concept, which cannot (or should not) be divided into subconcepts. A *compound descriptor* denotes a compound semantic concept, that is a semantic concept, which should be divided into its elements.

The approach of taking compound descriptors into the thesaurus is called *precoordination* in contrast to *postcoordination,* where a compound semantic concept is denoted by combination of simple descriptors during information retrieval. The distinction between precoordination and postcoordination is related to the decision whether descriptors are only single-word terms or, in general, compound words and multi-word phrases. Usually, a thesaurus is a compromise between both methods. There is a general rule for the decision where each one of the two methods should be used:

- *Postcoordination* should be used, when the morphological decomposition of a compound descriptor is identical with the semantic decomposition (*semantic factoring*). In other words, the logical sum of the semantic values of the components is equal to the semantic value of the compound descriptor. Otherwise the precision (see section 9) of information retrieval is decreased.

The advantages of postcoordination are in principle the advantages that have been mentioned in the discussion of the segmentation approach (see section 4).

Technical considerations may lead to variations of the given rule. For example, if there are two morphemes $m_1$ and $m_2$, which are used (nearly) always in the aggregate $m_1m_2$, than the compound descriptor may be taken into the thesaurus to save time, decrease ambiguity etc..

To some extent, ***relations*** can substitute precoordination. These relations can be fixed in the thesaurus and belong to the type of associative relations (see section 3.1 and 7.1). Or they can be payed regard to in the data structure (see section 7.2). It may be sufficient here to mention that a sophisticated data structure may replace a lot of precoordinated descriptors in medical language.

On the other hand, there is an important relation between ***dictionary size*** and the use of ***rewrite rules*** in the analysis of medical language data. This may become clear when considering a sample of synonymous phrases:

HERZINFARKT
CORONARINFARKT
INFARKT, CORONARGEFAESS
INFARKT, CORONARARTERIE
INFARKT, HERZKRANZGEFAESS
INFARKT EINER CORONARARTERIE

There are numerous ways of paraphrasing a given sentence without alteration of its meaning. The rewrite rules have to be handled with care, as an example may show, where each phrase represents another complex concept, constructed from the same two single concepts (CYST, INFLAMMATION):

| | | |
|---|---|---|
| CYSTITIS | - | CYSTITIS |
| INFLAMMED CYST | - | ENTZÜNDETE CYSTE |
| CYSTIC INFLAMMATION | - | CYSTISCHE ENTZÜNDUNG |
| INFLAMMATORY CYST | - | ENTZÜNDLICHE CYSTE |

## 7 Relations and Data Structure

### 7.1 RELATIONS

Relations have to be defined on every level of data representation. There are relations on the level of simple descriptors and segments as well as on the level of data aggregates. Some relations have already been mentioned and they are well known when classification and thesauri are discussed. Additional relations exist due to medical logic, for example,

- Antonyms (e. g., malignancy - benignancy),



- disease-symptoms (e. g., meningitis - rigidity of the neck),
- cause-effect (e. g., cytomegalic inclusion disease - cytomegalic inclusion disease virus; secondary effects of drugs),
- pathogenesis (e. g., obstruction *due to* tumor, paleness *due to* anemia, pneumonia ***complicating*** paralysis transverse),
- differential diagnosis - diagnostic procedure (e. g., meningitis - suboccipital puncture),
- location and location - specific alterations (e. g., limbus corneae - arcus senilis),
- influence of time with respect to the age of the patient and the age of the disease.

### 7.2 DATA STRUCTURE

A data structure is the physical basis for the formal representation of language data. It is not only a technical tool, but reflects also the selected semantic categories and additional features like time, origin of data, security, relations, sureness.

***Data atoms*** contain informations related to a single semantic category. ***Data aggregates*** can be built with varying levels and degrees of complexity: Signs, symptoms, syndromes, "simple" diseases, "complex" diseases.

There are some systems where the idea of the data structure is related to the concept of the ***Kernel sentence*** [13], which is a non-decomposible entity and carries the information of a sentence. It can be used to generate new sentences carrying the same information by application of well defined rules. Therefore the data structure cannot be seen separate from the structure of the thesaurus resp. classification underlying the thesaurus. In theory, the data structure has to reflect additional information like relations between Kernel sentences (***semantic operators***).

Besides logical problems, technical problems have to be considered to define a good data structure, which can be handled by man and/or by computer.

There are three systems for computer-based retrieval of medical language data utilizing the concept of the Kernel sentence for precise representation of information in a language data string. ACORN (Automatic Coding of Report Narrative) [3] is an automated natural language question-answering system for surgical reports. A Kernel extractor mappes the language data into the structure "F(x) = y", where x and y are "words or short phrases" and F is a function corresponding to phrases like "size of", "state of", "location" and so on. For example, a sentence like "the tumor is small" will be mapped into

SIZE (TUMOR) = SMALL .

The same will happen to all paraphrases of this sentence.

The authors point out, that natural language-based question-answering systems may even be simpler than systems based on artificial languages, which have to be learned by the human user. If we restrict relations between Kernel sentences to boolean relations, than answering a question formulated in natural language is nothing else than extracting the Kernel sentences out of the question and matching these sentences with the data base consisting of Kernel sentences.

Another system for automatic analysis of natural language pathology reports is based on the semantic categories: site, diagnosis and modifier [43]. The data are processed by a parsing routine, mapping into one out of five possible major data formats:

| Data format: | Example: |
|---|---|
| (1) DIAGNOSIS of SITE | Carcinoma of urinary bladder |
| (2) DIAGNOSIS-ive SITE (adjectival diagnosis) | normal vermiform appendix |
| (3) SITE-al DIAGNOSIS (adjectival site) | renal amyloidosis |
| (4) DIAGNOSIS = SITE-itis | Appendicitis |
| (5) SITE with DIAGNOSIS | uterus with leiomyoma |

The algorithm extracts delimiters out of the language string. These delimiters belong to a list of 62 morphemes, words or phrases, grouped into 13 groups, like conjunctions, diagnostic suffixes (-ITIS, -OSIS, ...), garbage words, diagnostic adjectival delimiters (-ating, -ed). After this, a preliminary assignment to one of the formats is done and tested for consistency. Inconsistency means, for example, that a diagnostic delimiter is found in a place, where the format expects a site delimiter. A limited set of rewrite rules is used (e. g., site adjectivals ⟶ site nouns, diagnostic nouns ⟶ site nouns). The authors report low error rates of about 10 % in 2 samples of diagnostic statements.

The third system will be explained in section 8.4.1 .

## 8 Systems for Management of Medical Language Data

### 8.1 SECTIONING OF REPORTS

In pure sectioning of reports there is no analysis of language data. Reports are formatted to a rigid scheme of sections and subsections and data input is done in natu-



ral language [40].

The pay-offs of this approach are:

- Higher degree of completeness of reports in a restricted domain of discourse, mostly connected to higher expenses in generating these reports,
- lower degree of ambiguity compared with the scanning of unformatted reports,
- increase in precision and improvement of communication,
- automated generation of reports and journals, automated monitoring of report generation if combined with man/machine systems (see section 8.3).

### 8.2 TECHNICAL SYSTEMS

In technical systems language data are handled mechanically. There is no data analysis except plausibility checking in terms of completeness and formal relations (e. g., mutually exclusive values of characteristics). Reports are generated by aggregation of selected items from various lists. Usually the items in a special list denote mutually exclusive informations. Some systems offer additional support by implementation of branches in the otherwise rigid sequence, dependant on simplistic text structure like "yes", "no", "male", "female" and so on [7, 22, 23, 28]. Sometimes supplementary remarks in free text are also possible.

There is a widespread use of technical systems in medicine with a great variety of data catching methods (mark sense forms, punched cards, CRT's) and flexibility. Some developments are combinations between purely technical systems and formulation of the input data in highly formalized artificial language [9, 12].

Technical systems offer possibilities, which may be extremely useful for further developments of medical information systems. The more specific the domain of discourse is, the more helpful these systems are. The main *advantages* are:

- *Sectioning* of reports,
- *standardization* of the "skeleton" of a report,
- combination between *documentation* and *report generation*, which ensures documentation.

The main *disadvantages* are:

- Need for usage of a special artificial language by physicians,
- high degree of formalization of input data resulting in a more time-consuming generation of reports.

## 8.3 MAN/MACHINE SYSTEMS

Man/machine systems are exclusively computer-based and require time-sharing and tele-processing facilities. Human intelligence is combined with the advantages of computers with respect to large storage capacity, fast access to data and fast processing of explicit procedures. Man/machine systems are in use for report generation [2, 17, 18, 21, 31], access to lexicons [8, 39], and augmenting the use of structured questionnaires.

## 8.4 DESCRIPTOR SYSTEMS

The replacement of a document by a set of descriptors is common to all these systems. It is intended, that the set of descriptors reflects the informational content of the document:

$$D \simeq \{d_1, d_2, \ldots, d_n\} \; .$$

In *descriptor-out-of-context* systems descriptors are added to the document, usually by human indexers after reading the document. In *descriptor-in-context* systems the descriptors have to be part of the language string from the document. Descriptor systems always depend on the existence of thesauri and dictionaries. The best known descriptor systems have been introduced in the documentation of literature (e. g., *Universal Decimal Classification*) and they are widely used in medicine.

The *International Classification of Diseases* (ICD) [33, 44] has emerged from a classification of causes of death presented in 1893. During a conference in 1900 delegates from 26 countries accepted the updated version of this classification of diseases and traumata. The current, 8th, revision has been published in 1967. There is an obligatory version encoded with 3 digits and an extended official version encoded with 4 digits. Based on the official ICD there are national adaptations like the American version ICDA and probably thousands of hospital- or even physician-based adaptations.

ICD is a logically ordered list of diseases, where each entry is denoted by a rigid numeric code. The numeric codes are running numbers. The main structure is by *topography* and on the second level by *nosology*.

The sections "malignant tumors" and "congenital malformations" have an inverted structure, e. g., the nosological main class "malignant tumors" is topographically structured on the second level.

In the dictionary sense the characteristics are:



- One-dimensional structure, mixed between nosology and topography,
- complete list of descriptors for the domain of discourse by definition,
- primitive semantic marker in the form of the numeric code,
- terminologic control of synonyms and quasi-synonyms.

To give better support for retrieval in the two dimensions topography and nosology, IMMICH [16] has developed the *Klinischer Diagnosenschlüssel* and a cross reference table to ICD. There is a 5 digit code equivalent with 2 digits for topography and 3 digits for nosology. The last nosological digit can be subdivided further for special needs.

The *Standard Nomenclature of Diseases and Operations* (SNDO) [34] is no longer maintained by its sponsors but some of the ideas have been absorbed in later developments.

The *Medical Subject Headings* (MeSH) [24] are used by human indexers in the MEDLARS-system to index the content of published medical literature.

Word oriented descriptor-in-context-systems are in widespread use. There are two better known systems following almost the same principles. One system is called the Lamson-IBM-Thesaurus [6, 19], the other one is the AGK-Thesaurus [32]. Both thesauri have been developed for pathology data processing and contain only single word descriptors (Fig. 18). The elements of the original text are called *input words*. In the first step these input words are replaced by *standard words*. The transformations performed during this step are:

- Replacement of *derivational forms* by their underived forms,
- replacement of *adjectival forms* by their noun forms,
- replacement of *synonyms* by their preferred terms,
- deletion of *insignificant* words (e. g., articles).

The second step contains:

- *Classification* of a standard word according to site, finding, and modifier,
- *subclassification* of site words and findings (one more level),
- *supply* of simple descriptors for compound descriptors, i. e., explicit notation of some hierarchical, partitive and associative relations (Fig. 19).

The AGK-thesaurus now contains about 50,000 input words and 12,000 standard words.

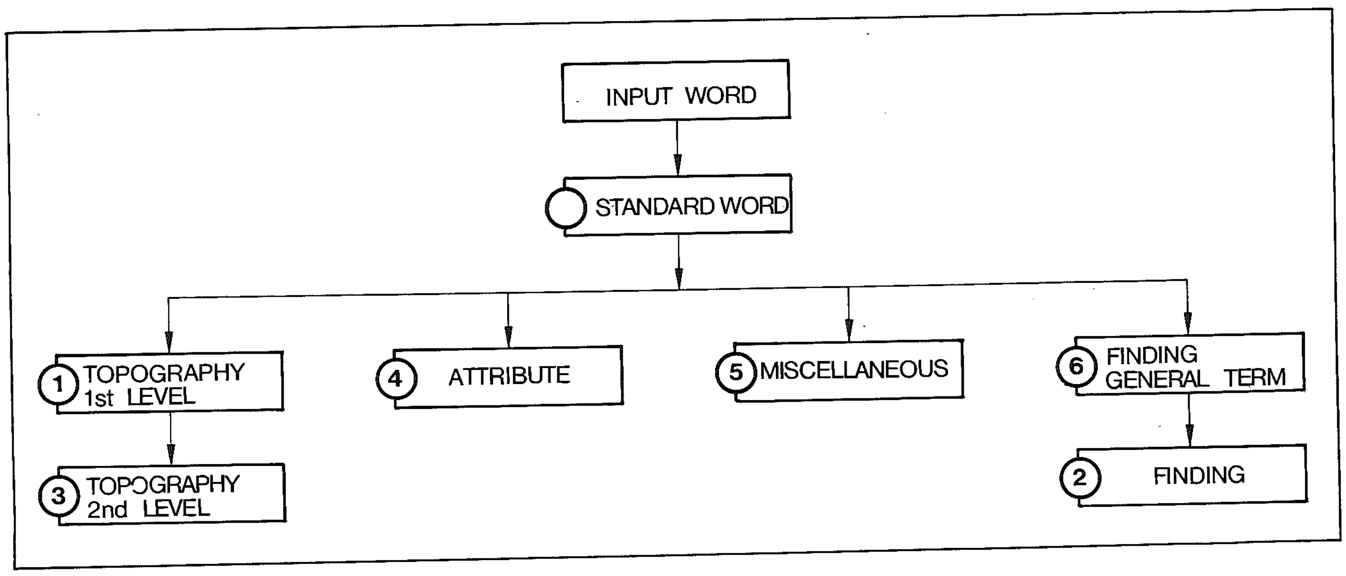



Fig. 18: Structure of the AGK-thesaurus [32].



| Input words | Standard word | Relations |
|---|---|---|
| NABELSCHNURPHLEGMONE<br>OMPHALOPHLEGMONE<br>UMBILICALPHLEGMONE | KEY: 4879 FAC. NR.: 2<br>NABELSCHNURPHLEGMONE | 1 ABDOMEN<br>1 INTEGUMENT<br>2 OMPHALITIS<br>2 PHLEGMONE<br>3 HAUT<br>3 NABEL<br>3 NABELSCHNUR<br>4 UMBILICAL<br>6 INFLAM/LOCAL |

Fig. 19: Example taken from the AGK-thesaurus [32].

Language processing is done by replacing every nonsignificant input word from a pathology report by its standard word and its explicit relations. Query is done by "standardizing" the search question in the same way and comparing the descriptor sets of the search question and of the documents. Some primitive syntactical informations can be used, like

- restriction of the data fields to single sentences,
- neighbourhood-relationships.

The main advantage of such a system is, that mainly the language practically used has to be considered and that there are almost no constraints on the pathologist. The main disadvantage is the amount of postcoordination that has to be done by the user during search phase if he wants to increase precision and recall (see section 9). Syntactical information is almost totally lost. Another disadvantage is the primitive data structure, which cannot serve as a basis for logical deduction.

In 1965 the College of American Pathologists published the *Systematized Nomenclature of Pathology* (SNOP) [38], which has been an important step forward. It provides the necessary thesaurus functions to describe diseases by a pathologist and it introduces an explicit data structure based on the semantic informations underlying the terms. SNOP contains about 15,000 single word- or multi-word-descriptors, which are categorized into 4 lists (Fig. 20). SNOP has been translated into German [42].

| | |
|---|---|
| Topography | - a list of the names of the body sites |
| Morphology | - a list of the names of structural changes that occur in tissues as a result of disease |
| Etiology | - a list of the names of causative agents of disease such as microorganisms, drugs and chemicals |
| Function | - a list of the names of the physiological manifestations associated with disease plus a limited number of specific infectious diseases |

Fig. 20: Semantic categories of SNOP [38].

There is a high degree of precoordination. Related terms are listed together, (quasi-) synonyms belong to the same equivalence class. Each list is highly structured, the depth of the structure varying between 2 and 3 levels. Each entry has a code-equiva-



lent consisting of 5 characters. The first character, T, M, E, F, denotes list membership, the following 4 characters are markers for the position of the descriptor in the respective list. Hierarchical relations, both generic relations and partitive relations are represented in the code structure (Fig. 21, 22). Code numbers on the same level are markers for the associative relation "parallel classes".

Another important step into the direction of a poly-dimensional approach is done by introduction of *cross references*. For example, the term APPENDICITIS is listed under M4001 and carries an additional semantic marker T66 for the site (APPENDIX). The cross references are only in the 2-digit form as pointers to topography.

The basic philosophy underlying the data structure is, that a diagnosis may be formulated by a pathologist as:

> There are *morphological* alterations in a *topographic* site due to an *etiologic* agent combined with a *functional* disorder.

Therefore, in general, all 4 lists have to be used in combination for full description of a pathology finding.

Handling of *modifiers* is done in SNOP in different ways. If a modifier is part of a concept, the compound descriptor is listed in the dictionary, either explicitly or implicitly. For example, in the inflammation section the modifier "acute" is used as a concept modifier and is encoded in the 2nd digit as number 1:

M4000 INFLAMMATION, NOS → M4100 ACUTE INFLAMMATION, NNB

M4000 CERVICITIS (T83) → M4100 ACUTE CERVICITIS (T83)

The same principle is used for "acute" when combined with leukemia, where it is encoded in the 4th digit as number 5:

M9803 LEUKEMIA, NOS → M9805 ACUTE LEUKEMIA

If the modifier "acute" appears in other concepts, it may be handled otherwise:

M3812 ACUTE CONGESTION, NOS

M5472 ACUTE INFARCT

The same modifier is not only used in concepts but also as a general modifier in me-

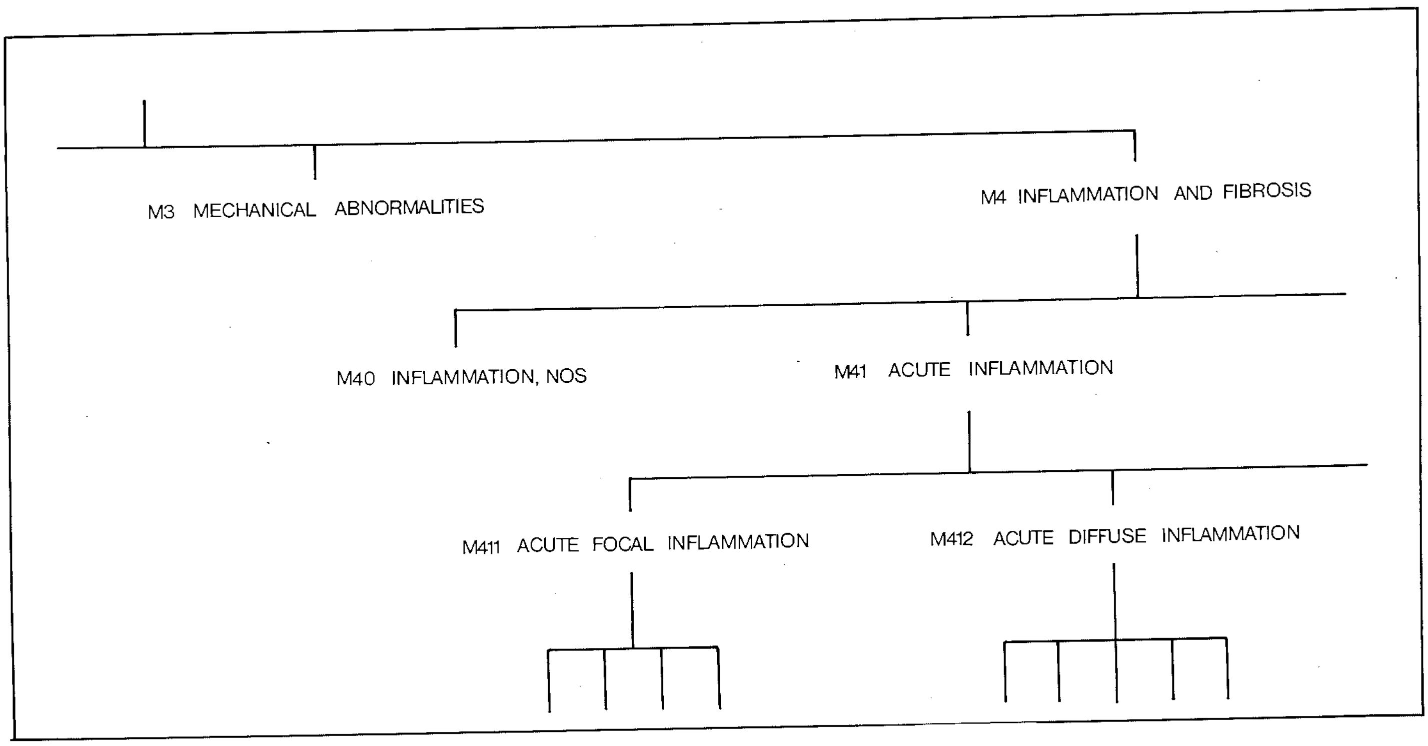



Fig. 21: Example of structure of SNOP (MORPHOLOGY) with generic hierarchical relations [38].



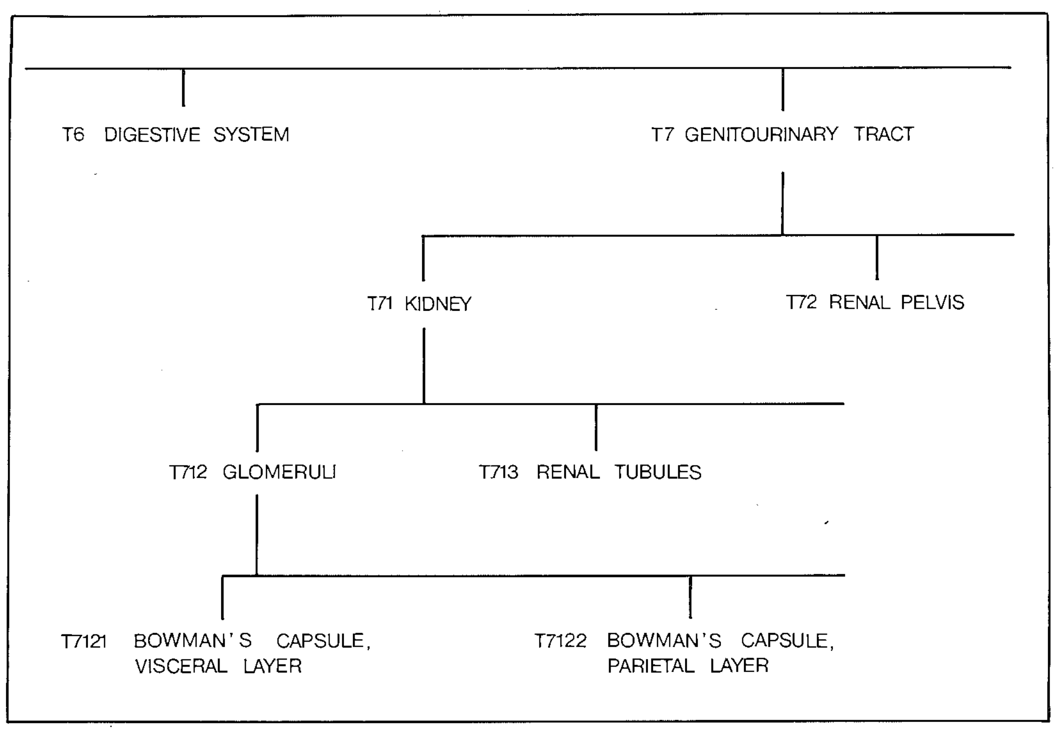



Fig. 22: Example of structure of SNOP (TOPOGRAPHY) with partitive hierarchical relations [38].

dicine, e. g., ACUTE ABDOMEN, ACUTE BLEEDING, and so on. Modifications like the latter ones cannot be coded, and they are lost if no special features are prepared for "general modifiers".

The success of SNOP led to the development of an expansion, called *Systematized Nomenclature of Medicine* (SNOMED) [5], which shall provide a dictionary for all of medicine. The major expansions are:

- Additional categories *procedures* and *complex diseases*.
- Increase in the number of descriptors (total number is about 40,000).
- Reclassification of several sections according to the development of medical knowledge (e. g., malignant tumors, leukemias, lymphomas). There is an important sign for convergence: the section "malignant tumors" of SNOMED will be identical to the analague section of the 9th revision of ICD (ICD-O).
- Expansion of the code to 6 digits in some sections.
- More specific cross references to topography.
- Expansion of cross references to other categories than topography.

The underlying assumption is, that a general utterance in medical language may be formulated as:

> There is a *procedure* for a *complex disease* or for *morphological* alterations in a *topographic* site due to an *etiologic* agent combined with a *functional* disorder.

There is one system for medical language data processing on the basis of SNOP developed by PRATT and his group [29, 30]. Each pathology finding is mapped onto the TMEF data structure and retrieval is done against this structure. All pathology data are stored in artificial language form and in their original form:

Txxxx Mxxxx Exxxx Fxxxx <Language string> .

The complete medical record therefore is a sequence of such TMEF statements (Fig. 23) (see section 8.4.1).

Another data structure is introduced in the REMAID-system, which can be arranged between technical systems and descriptor systems [15]. The physician has to use a special syntax and semantic categories: *symptoms, topography, etiology, treatment, results*. Descriptors have to be used according to dictionaries, which are different for different specialities. It is intended to implement the system for rheumatology



| | | | | |
|---|---|---|---|---|
| T2600 | M8103 | E0000 | F0000 | Bronchus, Carcinoma |
| T0000 | M0000 | E6927 | F0000 | Tobacco (Cigarettes) |
| T0000 | M0000 | E0000 | F7103 | Paroxysmal Nocturnal Dyspnea |
| T5600 | M8106 | E0000 | F0000 | Liver, Metastatic Carcinoma |
| T5600 | M3850 | E0000 | F0000 | Liver, Hemorrhage |
| T0000 | M7051 | E0000 | F0000 | Cachexia |
| T0000 | M0000 | E8816 | F0000 | Fluorouracil Therapy |
| . | . | . | . | |
| . | . | . | . | |
| . | . | . | . | |
| Txxxx | Mxxxx | Exxxx | Fxxxx | <Language String> |

Fig. 23: Example of a part of a medical record encoded due to SNOP [29, 30, 38].

### 8.4.1 A System for Pathology Data Processing

The most advanced system for automatic processing of medical language data has been developed by PRATT and his group [29].

For the explanation of the algorithm some definitions are necessary.

The formal structure of a dictionary entry of SNOP is represented in Fig. 24. SNOP is organized in indexed sequential manner, with the dictionary English $<k\ w_1 \dots w_m>$ as key. Access by the encoder program is done by specifying a character string v corresponding to a first word (k) of the dictionary English. By using the generic key feature, all entries having the string v as first word k are then available.

The encoding process is divided into 3 major steps:

(1) *Organization* of the original utterance into a special data structure, access to several dictionaries and associating of their contents with the utterance for processing in the following steps.

(2) Access to SNOP to obtain those dictionary entries relevant to encoding (*look-up matching* phase).

| | | |
|---|---|---|
| $\{k, W\}$ | = | <k $w_1 \dots w_m$ CR CODE > |
| k | = | first word of english language part (key) |
| $w_i$ | = | ith word of english language part, except key ($i = 1, 2, \dots, m$) |
| CR | = | cross reference SNOP-code |
| CODE | = | SNOP-code |

Examples:

| <k $w_1 \dots\dots\dots w_m$ | CR | CODE > |
|---|---|---|
| UTERUS LOWER SEGMENT | ∅ | T8230 |
| PHARYNX | ∅ | T6010 |
| PHARYNGITIS | T6010 | M4000 |
| INFLAMMATION BLENORRHAGIC | ∅ | M4040 |

Fig. 24: Formal structure of SNOP due to PRATT [29, 38].

(3) Use of dictionary entries obtained in step (2) and utterance structure obtained in step (1) to generate SNOP-statements (Fig. 23).

### 8.4.1.1 Preparation of an Utterance

From the original utterance an *item array* and an *item symbol array* are generated. *Items* are words or punctuation symbols. Abbreviations are regarded as words, for example, L4, meaning "4th lumbar vertebra".

Hyphens and quotation marks are deleted from the utterance. Abbreviations are expanded to their full form or to an abbreviation without a period. Some fixed phrases are transformed into single items ("consistent with" → CW).

Several dictionaries are accessed during this step. They are organized either as binary prefix trees in main storage or in indexed sequential manner on disc, depending on the frequencies of access and on available main storage space. Otherwise the dictionaries are equivalent. The dictionaries are accessed using as key the items or affixes obtained from the item array and deliver the elements of the item symbol array. One element of the item symbol array consists in general of different parts:



<item symbol> : = <mode> <type> <remainder>.

<mode> : identifies informations about the conditions under which the item symbol will be used. The values of <mode> may be:

- '0': obligatory transformation, exceptions to general rules, e. g., a word is used as key in nominal and in adjectival form,
- '/': item symbol is to be ignored,
- '-': the item has to be concatenated with its right neighbour,
- null: normal use of an item symbol.

<type> : indicates special functions to the encoder (Fig. 5), such as

- phrase delimiter,
- parts of speech (noun, adjective, ...),
- transformation from plural form to singular form for English, Latin and Greek words,
- transformation from adjectival to noun,
- transformation from noun to noun,
- deletion of some prefixes (MIDBACK → BACK),
- transformation of synonyms,
- replacement of plural forms of bilateral organs by their singular forms plus the words "LEFT" and "RIGHT",
- degree of relevance of a word with respect to SNOP,
- direct association of SNOP-codes (MUSCLE: T1300),
- specification of SNOP-code modifiers (ACUTE: CN21 CN45),
- exceptions to general rules.

The output of this step is an item array $u_i$ ($i=1,2,\dots,n$) and an item symbol array, which is a list of elements $t_{ij}$ denoting for every item $u_i$ possible transformations (<mode> = null in this example):

| $u_i$: | $t_{i1}$ | $t_{i2}$ | $t_{i3}$ |
|---|---|---|---|
| MALIGNANT | A | N1CY | |
| LYMPHOMA | N | | |
| TESTIS | N | A1CULAR | |
| BOWEL | NINTESTINE | | |
| ATYPICAL | A | N3X | N3A |

For explanation of the symbols used in the item array see section 4.2.1.

8.4.1.2 Look-up Matching Phase

There are two general parts in this step:

- *selection* of text structures from the original utterance,
- *transformation* of these text structures and *look-up* in the dictionary.

Words from the utterance may be used with two different functions. They may be used either as a *key* or as a member of a set of words related to another key (*match list*). The possible transformations $t_{ij}$ in the item symbol array are divided into two sets. One set of transformations is applied, if the word is used as key, and the other set of transformations is applied, if the word is a member of a match list.

A *minimal noun phrase* is a non-empty set of words in the context of a special word between words of type L (phrase delimiters including beginning and ending of the utterance).

From the item list a variable number of *syntactic potential noun phrases* is generated. A syntactic potential noun phrase consists of a key and of a match list:

$$\{v,V\} \quad = \quad \underbrace{t(v)}_{\text{key}} \quad \underbrace{v_1 \cdots\cdots\cdots\cdots v_n}_{\text{match list}}$$

All v's are words form the item array and t is a key-transformation of v. t may be the identity transformation. The elements $\{v,V\}$ are directly related to the SNOP-entries $\{k,W\}$. It is assumed, that the meaningful parts of the language, defined by the dictionary, may be paraphrased in the utterance in a way, that the semantic elements are contained in the syntactic potential noun phrases. Therefore, for look-up, the syntactic potential noun phrases have to be compared with the dictionary entries.

Selection of a Key

The item array is scanned and each word is marked as a possible key, if it has a letter N(noun) or A(adjective) as the first symbol in its item symbol list. Keys are selected from right to left. Within a minimal noun phrase first all nouns and then all adjectivals are selected.

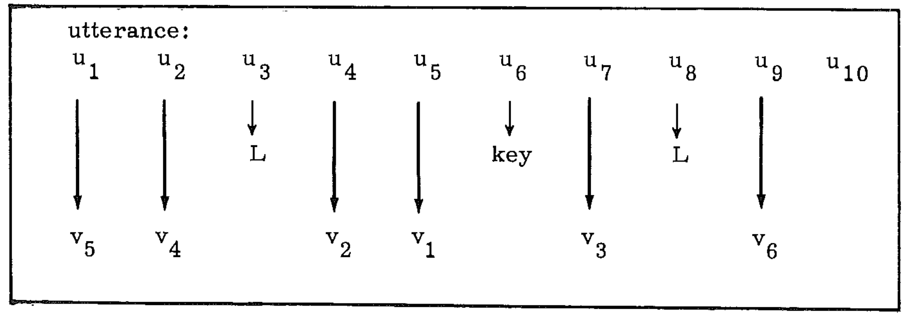


Selection of a Match List

To every selected key v a match list V is generated. Members of the match list are items surrounding the key in the item list. Selection is done using two different algorithms.

The first algorithm is used in case of *compound acjective phrases* such as "GASTRIC, PYLORIC, CELIAC, RIGHT COLIC AND AXILLARY LYMPH-NODES". For the noun-noun structure, "LYMPH-NODES", as key item, the match list will be in turn, each of the adjectivals or adjectival phrases to the left.

In all other cases each word to the left of the selected key and to the right of the next delimiter is selected. If there are less than 6 words, than each word to the right of the selected key and to the left of the next delimiter is selected. This process is repeated as long as there are less than 6 words in the match list and there are words left:




Key and match list represent, in general, a minimal noun phrase and fragments of other minimal noun phrases.

Look-up Phase

An element $\{k,W\}$ of SNOP is said to be matched with an element $\{v,V\}$ of the original utterance if the following conditions are met:

$$k = t(v)$$
$$\{w_1,\ldots,w_m\} \subset \{t_{ij}(v_i),\ j = 1,2,\ldots,n_i;\ i = 1,2,\ldots,n\}$$

That means:

- SNOP contains an element, whose key k is identical to a transformation t(v) of the selected key.
- Every additional word $w_1,\ldots,w_m$ is identical to a transformation of a member of the match list.
- Different words w have to be matched with different elements of the match list.

A selected key v is transformed for access to SNOP as long as there has been no match and there are possible key-transformations left.

After the successful use of a key item, the rightward extent of match lists for later key items may be restricted on the grounds that distinct dictionary entries from the same category are most likely parts of distinct noun phrases, and should not appear in the same match list. That is, inference about the syntactic structure is made from the semantic context. Autopsy summaries often contain long lists of topographic sites, separated by commas or semicolons. This technique limits the inclusion on match lists of significant words from distinct topographic sites occurring to the right of the key item in the utterance.

After a successful look-up on a key item, the SNOP-categories of the resulting dictionary entries are checked for consistency. If there is more than one SNOP-category represented, the restriction is not applied. A search to the right for the previous key item is made. If the results for this key item are also of the same consistent SNOP-category, and the first, or high order, digit of SNOP-code number of the last dictionary entries of each set of results do not match, it is assumed that the two key items belong to separate noun phrases of the same category:

| | |
|---|---|
| MALIGNANT | A N1CY |
| LYMPHOMA | N |
| ; | L |
| TESTIS | N OA1CULAR |
| , | L |
| CERVICAL | A * G |
| LYMPH-NODE | NLYMPH*%SNODE&LYMPH ANODE* |
| ,AND | L |
| TONSIL | N * |
| . | LB |
| T6110 TONSIL | |
| T0820 LYMPH NODE CERVICAL | |



After matching LYMPH NODE CERVICAL and comparing the results with those of the previous key, TONSIL, it is assumed that their respective noun phrases are unrelated. A "block" to halt the filling of match lists is placed at the previous key item. Under this modified condition, the program now selects a new key item and its match list. This simple conclusion on such general evidence is sometimes wrong. It does not appear to be possible to make accurate deductions about noun phrase structure in medical language on the basis of the semantic structure of the present SNOP-dictionary.

There are some more additional rules or exception rules, which will not be explained in detail.

For each selected key, pointers are generated to the items of the item list, which have been part of a match.

## Possible Improvements of Look-up Matching Phase

There is need for optimization of this step due to pragmatic and technical reasons. Minimal syntactical informations are involved in the generation of the syntactic potential noun phrases. After selecting the elements of a match list, no more syntactical information is used. This approach has two effects:

- Entirely different Kernel sentences will become undistinguishable:

| | | |
|---|---|---|
| NIERENBECKEN | | BECKENNIERE |
| KIDNEY PELVIS | - | PELVIC KIDNEY |
| T7200 | | T7100 M3315 |

- The step becomes very expensive. For each key-transformation t(v), up to $\prod_i n_i$ transformations in the match list have to be performed and compared with the set W. This product depends on the number of words in set V and it depends on the numbers $n_i$ of possible transformations for a word $v_i$. Another factor is the number of possible key transformations, which influences the number of disc accesses.

Because the whole algorithm is a match of a language string against a dictionary, improvements may be

- by identifying syntactic potential noun phrases with less "noise",
- applying only those transformations, which are highly productive.

Weaknesses of the dictionary result in longer lists of exceptions to general rules. In other words, weaknesses of the dictionary result in longer lists of possible transformations.

(1) Minimizing the Number of Syntactic Potential Noun Phrases

The number of syntactic potential noun phrases is determined by the number of keys. The number and types of keys are dependent on the structure of initial words of the dictionary. If the structure of the initial words can be described by symbols, which can be derived from an analysis of an utterance, than keys can be recognized and must not be guessed. (For example: keys may be only noun singular words).

(2) Reduction of the Number of Key-transformations

Other formal descriptions of the initial words may be developed, which reduce the number of key transformations to those, which are productive with respect to the dictionary. These transformations may be kept, for example, in a special dictionary based on morphemes or stems. Again, changes to the dictionary SNOP are alternatives to changes of the algorithm.

(3) Minimizing the Length of the Match List

An optimal solution would be a syntactic parser identifying the noun phrases. Experience shows, that general syntactic parsers are of no great value. In medical language some important features of natural language are rarely used, such as verbs, verb phrases, prepositions and so on. Words composing a noun phrase are frequently found in discontinuous sequences of the sentence.

The syntax of natural language may be violated while preserving the content carrying functions of the language (CORTICAL CYST OF THE KIDNEY).

(4) Minimizing the Number $\prod_i n_i$ of Transformations

This is done by minimizing the length of the match list (see (3)) and by minimizing the numbers $n_i$ (see (2)).

8.4.1.3 Mapping Onto the Data Structure

The procedures in this step serve the following purposes:

- Removal of dictionary entries carrying redundant or unwanted SNOP codes.
- Generation of the SNOP-statements combinatorially from the 4 categories represented in the dictionary entry array.
- Application of the SNOP-code number modifications for the inflammation, neoplasm, and leukemia code groups.
- Application of semantic consistency rules embodied in the code structure and cross-references of SNOP.
- Retentions or deletions for improving the readability of the encoder's output.



The respective algorithms are:

(1) Principle of longest match

This principle has already been mentioned. A SNOP-entry is deleted if his English language part is a subset of the English language part of another SNOP-entry, which has been matched. This principle is based on two contradictory assumptions:

- A SNOP-entry implicates every entry containing a subset of his English language part (hierarchical relations):

| |
|---|
| T0800 LYMPH NODE, NOS |
| T0820 CERVICAL LYMPH NODE, NOS |

| |
|---|
| M3850 HEMORRHAGE, NOS |
| M3851 HEMORRHAGE, PETECHIAL |

- A matched SNOP-entry consisting of an English language part, which is a subset of the English language part of another SNOP-entry, is an artefact:

| |
|---|
| [TOX40 PLASMA, NOS] |
| M9733 PLASMA CELL MYELOMA |

(2) SNOP-entries with 4 or more adjacent items dominate other SNOP-entries matching or overlapping on exactly one of these items and having at least 2 fewer matched items:

| |
|---|
| FOREIGN BODY GIANT CELL REACTION, KNEE JOINT |
| M4414 FOREIGN BODY GIANT CELL REACTION |
| T1272 KNEE JOINT |
| [M3911 JOINT BODY (T1200)] |

An exception is made when the overlapping item is a topography modifier such as LEFT.

(3) Expansion of the entry array by inclusion of the cross-references.

(4) Expansion of the entry array to include new entries generated from those unmatched items carrying a default SNOP-code in their item symbol list:

| | | |
|---|---|---|
| SMALL | | G |
| ARTERIES | | XT4100 ØSARTERY |
| OF | | L |
| LEFT | | GTL |
| ARM | | N |
| | | |
| TY802 | ARM LEFT | |
| T4100 | ARTERIES | |

(5) The dictionary entry array is reduced to include only one entry having any given SNOP-code. Duplicate codes are removed.

(6) SNOP topographic region codes (TY000 to TY999) are deleted, when the matched items overlap the matched items of a dictionary entry having a non-region topography code:

| | |
|---|---|
| T1111 | FRONTAL BONE |
| [TY011 | FRONTAL REGION] |

(7) The dictionary entry M9593 MALIGNANT LYMPHOMA is deleted, as redundant, if any of the malignant lymphoma class morphologies (SNOP codes M9603 to M9703, and M9750 to M9799) are present in the dictionary entry array:

| MALIGNANT LYMPHOMA, HODGKIN'S GRANULOMA | |
|---|---|
| [M9593 | MALIGNANT LYMPHOMA, NOS] |
| M9673 | HODGKIN'S GRANULOMA |

(8) Dictionary entries carrying the general neoplasm class codes (M8000 to M8009) are deleted, as redundant, from the dictionary entry array when a more specific neoplasm class code (M8010 to M9799) is present.

(9) SNOP-code modifications (see section 8.4) are applied to appropriate codes, if and only if there is exactly one dictionary entry in the particular code modification class. The item array is scanned from left to right for items, such as ACUTE, which indicate code modifications but which are unmatched. Modifications to a SNOP-code number digit are not made if the digit already has been modified.

(10) SNOP-statements are now generated combinatorially from the remaining dictionary entries. Some of these are deleted because of inconsistency of SNOP-codes in the generated statement with the cross-reference SNOP-codes carried by the dictionary entry of another category of the statement.

To generate a SNOP-statement, each topography code in the dictionary entry array is paired with each morphology code and each of these pairs with each etiology code, then each of these triples with each function code.

Those SNOP-statements are not retained if a cross-reference SNOP code carried by one of the dictionary entries is inconsistent in the first two digits of SNOP-number with the SNOP-code of the same category in



CERVICITIS, CHRONIC CYSTIC AND ATROPHIC ENDOMETRIUM

| | | |
|---|---|---|
| M4384 | CERVICITIS, CHRONIC CYSTIC (T83) | matched entries |
| M7837 | ATROPHIC ENDOMETRIUM (T84) | |
| T8400 | ENDOMETRIUM | |
| T8300 | CERVIX UTERI | |

| | | | | |
|---|---|---|---|---|
| T8300 | M4384 | E0000 | F0000 | SNOP-statements |
| T8300 | M7837 | E0000 | F0000 | |
| T8400 | M4384 | E0000 | F0000 | |
| T8400 | M7837 | E0000 | F0000 | |

The combinatorial generation of SNOP-statements from an array of dictionary entries assumes, that the utterance for which the statements are generated, states that all members of each SNOP-category are present with each member of the other SNOP-categories. This is frequently true. For example, in surgical pathology an examination of a single specimen is fairly common, or a single diagnosis from the morphology category may be stated for several specimens from the same case. It is also frequently untrue, for example, "necrosis of spleen, hyperemia of liver". The algorithm uses the semantic rules implied by the cross-reference SNOP-codes, where available, to purge some of the possible erroneous combinations. However, the errors arising from this method cannot be generally eliminated without the use of medical language syntax to help establish the proper semantic combinations to be generated from dictionary entries matched on noun phrases.

In this final phase a sequence of individually simple algorithms, some of them semantically contentless, has been used for a kind of semantic processing. When a dictionary is developed with a more complete representational structure for its domain of medicine the linguistic processes operating on a well-developed semantic dictionary should become both simpler and more interesting academically.

The viability of the current techniques against the current dictionary may be summed up as follows. A single simple algorithm can often make multiple changes to a data structure. These changes may cut across the boundaries of quite diverse portions of a conventional, intuitively satisfying linguistic or semantic theory. The "meaning" of these contentless algorithms is partly to be found in their sequential relationships. In constructing them, one considers what structure the data have before and after their application, which algorithm follows, and if the effect produced at this point in the sequence is desirable. This kind of system is relatively easy to program but the interdependencies among its components make certain kinds of modifications difficult or impossible.

Modifications, which do fit easily into this framework are treatment of high frequency problems, which can be recognized from SNOP-codes, such as the treatment of the malignant lymphomas and topographic regions.

### 8.4.1.4 Adaptability of the Algorithms to German Medical Language

So far, there is no reason why the algorithms could not be applied to German medical language too. Our experience is, that there are neither serious semantical nor structural problems. Nevertheless, the direct implementation would be inefficient because of the widely used possibility of generating compound words in German language.

Listing all the possible paraphrases in SNOP, would probably lead to an enlargement by a factor of about 4. Furthermore,most compound words denote compound concepts from several semantic categories. This would result in a much more complex cross-reference structure. For this reason a modification using the segmentation approach should be done. If the basic entity "word" is replaced by "segment", than SNOP can even be reduced.

Therefore, semantic markers have been associated with segments. Each segment, which appears only in a single SNOP-entry, is marked by the respective SNOP-code. Each entry in SNOP, whose semantic content can be generated from the semantic contents of its segments can be deleted from SNOP. In other words, if the segmentation of SNOP-entries delivers the same result as the semantic factoring does, then the SNOP-entry can be generated both by language and by meaning.

### 8.4.1.5 Expansion of the Data Structure

SNOMED [5] is the basis for expansion of the above described methods to medicine. It imposes a fundamental data structure, which has to be proved in practical use.

Depending on the accurateness of encoding, provision has to be made to encode modifiers, which are not part of the concepts listed in SNOMED (e. g., size, degree, time, number, quality, extent, and so on). Furthermore relations have to be established between the fundamental SNOMED-statements (Fig. 25).

The language tokens and the grammatical functions for representation of these relationships have to be defined. This is a medical and a linguistic problem. After this ... it seems to be possible to develop algorithms for the recognition of



- pathogenesis

  TUMOR OBSTRUCTING SMALL INTESTINE

  EPIDURAL HEMATOMA COMPRESSING TEMPORAL LOBE

- logical relationships

  BLINDNESS DUE TO ACCIDENT

- qualifications

  CARCINOMA OF THE ADRENAL GLAND WITH EXTENSION INTO THE RENAL CORTEX

  CARCINOMA OF THE STOMACH INFILTRATING INTO THE PANCREAS

- relations between elements of PTMEF-statements

$$B\left(A\left\{I\left[L(t_i,\ m_j),\ e_k\right],\ f_\ell\right\},\ p_n\right)$$

A statement like this one would be read as:

a procedure $p_n$ "B" ("as therapy for") a lesion $m_j$ L ("in", "contained in", "of") site $t_i$ I ("due to") $e_k$ A ("associated with") $f_\ell$.

Fig. 25: Examples for relations which have to be considered in medical language data processing using SNOMED [5].

the full semantic content of language data and for data structures, which allow for formal representation and retrieval of the semantic content.

## 9 Automated Retrieval of Language Data

Automated retrieval of language data is done by

- document identification (patient identification),
- content.

Retrieval by document identification shall not be discussed here, because it offers no special problems in medical language data processing. Retrieval by content means, that there is a search question, asking for all documents whose semantic content matches the semantic content of the question.

If there is only a *sectioning of reports* (see section 8.1), then information retrieval is accomplished in one of two ways:

(1) Retrieval of the content of (sub-)sections,

(2) scanning of specified (sub-)sections for the appearance of user-defined descriptors.

In *technical systems* (see section 8.2) data search may be done by explicit data field and value. In computer-based systems boolean expressions are used for query across data fields. Additional specification of strings for free text supplements may be possible. The same remarks can be done for *man/machine systems* (see section 8.3).

In *descriptor systems* (see section 8.4) data search is done by explicit or implicit specification of descriptors. This is most easily be understood in descriptor-in--context-systems, using word-oriented thesauri. For every descriptor asked for, all its synonyms and all the descriptors in a downward hierarchical relation are additionally supplied. This has to be done either by man or by computer.

In systems using the idea of Kernel sentences, search questions are manipulated in the same way as language data. The Kernel sentences are extracted from the search question and these Kernels are compared with the data base of Kernel sentences from the original data, if there is no cross-referencing between different Kernel sentences.

Measures for the quality of retrieval systems

The best known measures for the quality of retrieval systems are

$$\textit{recall} = \frac{\text{number of relevant elements retrieved}}{\text{number of relevant elements in the collection}} ,$$

$$\textit{precision} = \frac{\text{number of relevant elements retrieved}}{\text{number of elements retrieved}} .$$

The value of the two measures ranges between 0 and 1. Ideally, both values are equal to 1. The computation of the two measures requires a knowledge not only of what is retrieved, but also of what is relevant to a given query. For the *recall*-computation,



furthermore, information must be available concerning the number of relevant items in the collection. This information is available only for small test queries because it requires additional indexing by experts. In general, the two measures are dependant, and improvement of one measure results in deterioriation of the other one.

Another problem is the definition of *relevance*. If we use descriptor systems, than we can represent the documents $D_i$ in the collection in a first level approach by a set of descriptors

$$D_i \simeq \{d_{i1}, d_{i2}, \ldots, d_{im_i}\} \quad , \quad i=1,2,\ldots,m ,$$

and the query q by another set of descriptors

$$q = \{q_1, q_2, \ldots, q_n\} \text{ with the notion of length of } q = n .$$

It is fair to say, that a document $D_i$ is relevant, if q is a subset of $\{d_{ij}\}$. Looking at normal communication, a well defined boundary between relevant documents and irrelevant documents is unnatural. Medical informations build an information continuum and discrete dots are enforced only when people communicate. It may be possible to find a fuzzy definition of "relevance" in terms of:

> a document is *relevant*, if *some* or *most* of the descriptors $q_i$ are members of the set $\{d_{ij}\}$ .

Then we may give the following definition of

$$\textit{degree of relevance} = \frac{\text{number of descriptors } q_i \text{ matched with set } \{d_{ij}\}}{n} .$$

In this case it may be up to the user to define his own boundary of relevance.

Using this approach, every descriptor has the same weight. This may be changed for descriptor systems with different information "load". For example, in a medical descriptor system, there is not much information in descriptors like "lobe", "disease", "morbus", "disturbance", and so on. More information is contained in descriptors like "heart", "carcinoma", "right upper lobe", and so on. Therefore, a weight may be introduced for each descriptor in the system. This may be done by experts, or it may be done by a formal algorithm. A better known formal algorithm is a definition of the weight of every descriptor proportional to a number 1/p, where p is the relative frequency of the descriptor in the data base. This introduces the notion of a *rank number* R for every document with respect to a given query:

$$R \sim \sum_{k=1}^{n} \delta_k / p_k \,, \qquad \delta_k = \begin{cases} 0, & q_k \notin \{d_{ij}\} \\ 1, & q_k \in \{d_{ij}\} \end{cases}$$

If a thesaurus is used, another approach may be taken to define some sort of neighbourhood to a special dot defined by a query. ***Associative relations*** may be used to define for each pair of descriptors the degree S of similarity between the semantic concepts denoted by the descriptors. If this is the case, the query may be extended from each descriptor $q_i$ to a set of descriptors $r_j$ with $S(q_i, r_j) \geq p$ where p has to be selected by the user. This type of similarity is defined implicitly in a nomenclature like SNOP or SNOMED by the neighbourhood in the respective lists and it has become explicit in the similarity of the SNOP-codes. Due to the hierarchical structure of the SNOP-code, a "distance" between two semantic concepts may be defined in terms of the difference of the respective SNOP-codes. It is quite clear, that the code-equivalent of a descriptor is not a value of a quantitative variable called T or M. Nevertheless the hierarchical organization introduces some properties of mathematical variables. It is possible to define in the M-category a subspace by the limits M4 and M4999(inflammation), or M41 and M4199 (acute inflammation) or M411 and M4119 (acute focal inflammation). Using this notion of a variable, reflecting the similarity of medical concepts, queries like the following one seem to be naturally formalizable:

> All documents with inflammatory diseases of bones, caused by bacteria and associated with pain:
>
> T11** M4*** E1*** F753*

This query defines not only a set of dots but also a 4-dimensional subspace of the total information space defined by SNOP. These features have been used for the development of a retrieval language [29]:

> All records of "primary choriocarcinoma of the uterus with metastases to the liver and with no metastases to the lung":
>
> IF SNOP EQ T8300 M8823 /
> AND SNOP EQ T5600 M8826 /
> AND NOT SNOP EQ T2800 M8826.



## 9.1 DATA PRESENTATION

There are other improvements that are enabled in data presentation using a semantically structured lexicon that may only be demonstrated shortly. If we look at an autopsy report represented as a set of dots in the 4-dimensional discrete information space defined by SNOP, it is possible to give formalized projections of this report using 2 dimensions:

| T \ M | 2919 | 3850 | 4046 | 4060 | 4200 | 5453 | 7634 | 8806 | 8809 |
|---|---|---|---|---|---|---|---|---|---|
| 0470 | | | | | | | * | | |
| 0X00 | * | | | | | | | | * |
| 2800 | | | * | | | | | | * |
| 3100 | | | | * | | | | | * |
| 5600 | | | | | | | | | * |
| 5610 | | | | | * | | | | |
| 7110 | * | | | | | | | | |
| X240 | | | | | | * | | | |
| Y240 | | | | | | | | * | |

It is quite clear that a data presentation like this one provides some capabilities of pictures. A whole set of informations can be seen at "one glance". The same type of data presentation may be used for the projection of a 3rd dimension (write the 3rd "coordinate" in the matrix element) or for a statistical analysis (write the frequency in the matrix element). Matrixes like the one demonstrated may be further input to statistical analysis for classification algorithms.

## 10 Literature

[1] ANDERSON, J.: Information System in the Hospital
Lecture Notes

[2] BARNETT, G.O., GREENES, R.A., GROSSMAN, J.H.: Computer Processing of Medical Text Information
Meth. Inform. Med. 8 (1969) 177 - 182

[3] BROSS, I.D.J., PRIORE, R.L., SHAPIRO, P.A., STERMOLE, D.E., ANDERSON, B.B.: Feasibility of Automated Information Systems in the Users' Natural Languages
Americ. Scientist 57 (1969) 193 - 205

[4] CHOMSKY, N.: Aspects of the Theory of Syntax
Cambridge, Mass.: MIT-Press 1965

[5] COLLEGE OF AMERICAN PATHOLOGISTS: Systematized Nomenclature of Medicine. Trial Version
Chicago: College of American Pathologists 1975

[6] DIMSDALE, B.: A Thesaurus Processor
IBM Form 320 - 2628 (1969)

[7] VAN EGMONT, J., WIEME, R.J.: Systematized Codification of Medical Diagnostic Statements
In: ANDERSON, J., FORSYTHE, J.M. (eds.): Proceedings of the MEDINFO 74
Amsterdam: North-Holland Publ. Comp. 1974

[8] ENGELBRECHT, R., SCHERTLEIN, G., REICHERTZ, P.L.: AKOS - Allgemeines Codierungssystem
In: REICHERTZ, P.L. (Hrsg.): Das medizinische System Hannover (MSH)
IBM Form G 12 - 1166 (1972)

[9] GIERE, W., BAUMANN, H.: Zur Erfassung und Verarbeitung medizinischer Daten mittels Computer. 1. Mitteilung: Ein Datenerfassungs- und Speicherprogramm (DUSP) zur Dokumentation von Krankengeschichten
Meth. Inform. Med. 8 (1969) 11 - 18

[10] GOOS, G.: Basic Notions of Informatics
Lecture Notes



[11] GORDON, B.L. (ed.): Current Medical Information and Terminology (CMIT)
Chicago, Ill.: Americ. Medic. Ass. 1971

[12] HALL, P., MEILNER, CH., DANIELSON, T.: J5 - A Data Processing System for Medical Information
Meth. Inform. Med. 6 (1967) 1 - 6

[13] HARRIS, Z.S.: Mathematical Structure of Language
New York: Wiley 1968

[14] HARTMANN, F.: What Could and Should Doctors Learn From Their Experiences With Computers in Medicine
In: ANDERSON, J., FORSYTHE, J.M. (eds.): Proceedings of the MEDINFO 74
Amsterdam: North-Holland Publ. Comp. 1974

[15] DE HEAULME, M., MERY, C.: REMAID: An Artificial Language for Medical Reports on Computer
In: ANDERSON, J., FORSYTHE, J.M. (eds.): Proceedings of the MEDINFO 74
Amsterdam: North-Holland Publ. Comp. 1974

[16] IMMICH, H.: Klinischer Diagnosenschlüssel (KDS)
Stuttgart: Schattauer 1969

[17] KOEPPE, P.: Zum Problem der EDV-gerechten Erfassung medizinischer Befunde
Meth. Inform. Med. 10 (1971) 25 - 29

[18] KOEPPE, P., SCHAEFER, P., GUTENMORGEN, W., SCHWOERE, I.: Das System ORVID, ein Beitrag zur programmierten Dokumentation in der Röntgendiagnostik
Fortschr. Röntgenstr. Nukl. Med. 112 (1970) 103 - 110

[19] LAMSON, B.G.: Natural Language Retrieval System, Pathology Thesaurus
UCLA 1970

[20] LEIBER, B.: The German Syndrome Identification and Information System (DOFONOS)
Meth. Inform. Med. 14 (1975) 69 - 72

[21] LODWICK, G.S., REICHERTZ, P.L., PAQUET, E., HALL, D.L.: "ODARS", A Computer Aided System for Diagnosing and Reporting
Part I: Clinical Problems
In: DE HAENE, R., WAMBERSIE, A. (eds.): Computers in Radiology

(Proceedings of the International Meeting on the Use of Computers in Radiology, Brussels, Sept. 1969)
Basel (1970)

[22] MAYNE, J.G., WEKSEL, W.: Toward Automating the Medical History
Mayo Clin. Proc. 43 (1968) 1 - 25

[23] MAYNE, J.G., MARTIN, M.J., MORROW, G.W., TURNER, R.M., HISEY, B.L.: A Health Questionnaire Based on Paper-and-pencil Medium Individualized and Produced by Computer
J. Amer. Med. Ass. 208 (1969) 2060 - 2068

[24] NATIONAL INSTITUTES OF HEALTH: Medical Subject Headings, Index Medicus
NIH Publ. No. 72 - 265, Vol. 13, part 2
Nat. Libr. of Med. U.S. Gov. Printing Office, Washington, D.C.

[25] PACAK, M.: Computational Linguistics and Information Handling
In: HOWERTON, P.W. (ed.): Management of Information Handling Systems
Rochelle Park, New Jersey: Hayden Book Comp. Inc. 1974, 19 - 47

[26] PACAK, M.: Automated Morphosemantic Analysis of Compound Word Forms in Medical Language
Personal note

[27] PACAK, M., COUSINEAU, L., WHITE, W.: The Segmentation Approach to Dictionary Construction
National Institutes of Health: NIH-Publ. 32801

[28] POCKLINGTON, P.L.: AMAP - A General Optical Mark Reader Form Evaluation Program
Meth. Inform. Med. 12 (1973) 211 - 222

[29] PRATT, A.W.: Medicine, Computers and Linguistics
Adv. Biomed. Engineering 3 (1973) 97 - 140

[30] PRATT, A.W., PACAK, M.: Identification and Transformation of Terminal Morphemes in Medical English
Meth. Inform. Med. 8 (1969) 84 - 90

[31] REICHERTZ, P.L., LODWICK, G.S., PAQUET, E., HALL, D.L.: "ODARS", A Computer Aided System For Diagnosing and Reporting.
Part I: Technical Problems



In: DE HAENE, R., WAMBERSIE, A. (eds.): Computers in Radiology (Proceedings of the International Meeting on the Use of Computers in Radiology, Brussels, Sept. 1969)
Basel (1970)

[32] RÖTTGER, P., WINGERT, F., FEIGL, W., GRAEPEL, P., GROSS, W.M., RIES, P., MATAKAS, F.: Structure and Development of a Thesaurus for Accomodation of Autopsy and Biopsy Records to Automatic Free Text Evaluation
4th Congress of the "Europäische Gesellschaft für Pathologie", Budapest 21. 9. 1973

[33] STATISTISCHES BUNDESAMT: Internationale Klassifikation der Krankheiten (ICD), 8. Revision
Stuttgart, Mainz: Kohlhammer 1968

[34] THOMPSON, E.T., HAYDEN, A.D. (eds.): Standard Nomenclature of Diseases and Operations
New York: McGraw-Hill 1961

[35] UICC (International Union Against Cancer): Illustrated Tumor Nomenclature
Berlin, Heidelberg, New York: Springer 1969

[36] UICC (International Union Against Cancer): TNM Classification of Malignant Tumors
Geneva 1974

[37] UICC (International Union Against Cancer): TNM General Rules
Geneva 1974

[38] WELLS, A.H. (chm.): Systematized Nomenclature of Pathology
Committee on Nomenclature and Classification of Disease
Chicago, Ill.: College of American Pathologists 1965

[39] WINGERT, F.: Klartextverarbeitung in der Pathologie
In: REICHERTZ, P.L. (Hrsg.): Das medizinische System Hannover (MSH)
IBM Form G12-1166 (1972)

[40] WINGERT, F.: Pathologie-Befund-System
Meth. Inform. Med. 12 (1973) 150 - 155

[41] WINGERT, F.: Word Segmentation and Morpheme Dictionary for Pathology Data Processing
In: ANDERSON, J., FORSYTHE, J.M. (eds.): Proceedings of the MEDINFO 74
Amsterdam: North-Holland Publ. Comp. 1974

[42] WINGERT, F.; GRAEPEL, P.: Systematized Nomenclature of Pathology, Deutsche Übersetzung
Schriftenreihe des Instituts f. Mediz. Informatik und Biomathematik der Univ. Münster 1975

[43] WONG, R.L., GAYNON, P.: An Automated Parsing Routine for Diagnostic Statements of Surgical Pathology Reports
Meth. Inform. Med. 10 (1971) 168 - 175

[44] WORLD HEALTH ORGANIZATION: International Classification of Diseases, 8th Revision (ICD/VIII)
Geneva 1967

# The computer utility in a health enviroment

**W. Schneider**

## 1. Introduction

A health information system (HIS)must be arranged so as to be useful for widely different circumstances of care. These include information processing where requirements have different timing requirements, levels of detail, need for readiness of access, etc.

a. For anticipatory screening to prevent illness and disability and to achieve improved personal health as well as to recognize and control early disease resk factors.
b. For periodic and episodic acute or short term care designed to remedy or cure illnesses and injuries. This may involve management of life threatening situations. Data access is almost unpredictable in respect to whom and when.
c. For comprehensive, continuing, or extended care (often called "long term"), where the purpose is to restore healthful status or to achieve rehabilitation or restoration of function. Predictability of usage is high and repetitive. Extended life maintaining services may have to be aided when the original medical condition is incurable or unavoidable as can occur with advanced age, etc. In this circumstance it is crucial that the system assist in achieving comprehensiveness and continuity.
d. For population health educational purpose largely directed toward better personal health practices preserving the capacity to cope with illness and disability. Here, generality of data usage and elective usage of data is common.

It also becomes evident that the most promising strategy for further development of computer usage in health care is to create a common data base (CDB) as a core around which the various - step-by-step developed - application systems are grouped. Together with the common data base these application systems form the HIS. They can be considered as subsystems of HIS, and it is this notion which is used in this report.

There are mainly two types of such HIS-subsystems. One type, which may be called CDB independent or dedicated, is used to improve information handling and communication within a structural element of a health