Susanne Wolff

The Use of Morphosemantic Regularities in the Medical Vocabulary for Automatic Lexical Coding

(From the Linguistic String Project, New York University)

Summary

The Neo-Latin portion of medical vocabulary exhibits morphosemantic regularities that make it possible to determine both English syntactic subclasses and medical semantic subclasses from formal properties of these lexical items alone. This paper describes an experimental program for the automatic creation of dictionary entries that exploits the formal regularities to obtain dictionary entries sufficient for computerized medical text analysis as presently carried out by the New York University Linguistic String Project (LSP) system. Although automatic dictionary preparation does not supersede manual classification, the program takes a considerable burden off the dictionary worker's shoulders and speeds the costly preprocessing stages of computerized text analysis.

Key-Words: Medical Terminology, Morphosemantic Analysis, Automated Dictionary Classification

Die Verwendung von morphosemantischen Gesetzmäßigkeiten im medizinischen Vokabularium für automatische lexikalische Kodierung

Der neo-lateinische Teil des medizinischen Vokabulars zeigt morphosemantische Gesetzmäßigkeiten, die es ermöglichen, sowohl englische syntaktische Untergruppen als auch medizinische semantische Untergruppen lediglich aufgrund der formalen Eigenschaften dieser lexikalischen Sachverhalte zu bestimmen.
In dieser Arbeit wird ein experimentelles Programm zur automatischen Herstellung von Wörterbucheintragungen beschrieben, welches die formalen Gesetzmäßigkeiten ausnutzt, um Wörterbucheintragungen zu erhalten, die für die automatisierte medizinische Textanalyse, wie sie derzeit durch das Linguistic String Project (LSP)-System der New York University durchgeführt wird, ausreichen. Obwohl die automatische Vorbereitung von Wörterbüchern die manuelle Klassifikation nicht ersetzt, entlastet das Programm die an einem Wörterbuch Arbeitenden erheblich und beschleunigt die teuren Vorstadien der automatisierten Textanalyse.

Schlüssel-Wörter: Medizinische Terminologie, morphosemantische Analyse, automatisierte Wörterbuchklassifikation

Introduction

The natural language information processing system developed by the Linguistic String Project (LSP) at New York University converts the information in narrative patient data into structured formats so that it can be accessed by computer programs for retrieval, summarization, and other informational tasks [5, 14, 15].

A dictionary containing entries for all text words is an integral part of this system. Dictionary entries specify the lexical category of a word, i.e. they identify it as a noun, adjective, or any of the other word classes represented by a terminal symbol of the grammar. Additional syntactic and semantic properties of the word are specified in the form of attributes (also called *subclasses*). During text processing the categories and subclasses of cooccurring text words are tested for syntactic and semantic compatibility in order to rule out incorrect parses and to arrive at an informational characterization of the input sentences [3].

The automated linguistic analysis of large bodies of natural language data requires substantial, and, as past experience with the LSP processor has shown, fairly detailed dictionaries. Although large dictionaries in machine-readable form do exist—e.g., Webster's Seventh Collegiate Dictionary [17] or Longman Dictionary of Contemporary English [7]—they often provide little more than the citation forms, rudimentary part-of-speech information, and definitions in a form unsuited for computational purposes. The LSP system has therefore relied on the manual preparation of its dictionaries. In a preprocessing stage all text words are coded for their major syntactic category membership and as thoroughly as possible for the additional syntactic and semantic properties used by the text processing program.

This preprocessing stage is time-consuming in that it requires of the dictionary workers that they operate within the grammatical framework of the system; i.e. it presupposes extensive familiarity with the ca. 150 classes and subclasses relevant to the LSP English language processor (defined in [2]). As a result, the computer dictionaries of the LSP have grown slowly. The English dictionary currently in use at the LSP contains approximately 10,000 entries, covering only a small portion of the English word stock when compared with Webster's Seventh New Collegiate Dictionary (ca. 100,000 entries) [17] or Webster's Third International Dictionary (500,000 entries) [18] . While the LSP English dictionary can be supplemented by one of the machine-readable large dictionaries to achieve satisfactory English parsing [19, 20], the vocabulary and classification provided by these sources is not adequate for processing the information in documents in specialized areas (i.e. documents of a "sublanguage"). For

this task, the LSP English dictionary is complemented by various specialized dictionaries whose entries are coded for the semantic subclasses that express the informationally significant categories of the subject area in addition to the English classes and subclasses. The assignment of sublanguage semantic subclasses to text words not yet found in the existing dictionaries of the LSP places a further burden on the dictionary worker who cannot be expected to be a specialist in the field.

Under ideal circumstances, the dictionary accessed by the language processor contains all the text words, or almost all words, so that preprocessing is kept at a minimum. Practice falls short of this ideal. Our work on medical documents has shown that approximately a third of the text words of incoming documents are not listed in the LSP medical dictionary of (currently) 6,500 entries so that this dictionary has to be constantly updated. While the proportion of words not found versus words already listed in the medical dictionary decreases with every set of documents to be processed, the successful machine analysis of large bodies of clinical narrative requires a much more substantial dictionary than presently available. Other than the vocabulary shared with general English, a sublanguage dictionary for this subject area should contain entries for the ca. 100,000 words in medical dictionaries such as Dorland's Illustrated Medical Dictionary [1] or Stedman's Medical Dictionary [16].

To remedy this situation, we have investigated whether any portion of the vocabulary of the sublanguage of clinical reporting lends itself to an automatic analysis and classification that is both detailed and reliable enough to serve as a viable alternative to manual dictionary preparation. A prerequisite for computerized dictionary preparation is the existence of words whose lexical category membership, English and sublanguage-specific subclasses can be inferred from formal properties alone. These formal properties, furthermore, have to be statable on the graphemic level in order to be recognized by the computer.

The research reported here has been conducted over a three-year period, from 1980 to the present. The work was stimulated by published findings of Pacak, Norton, and Dunham on the morphosemantic analysis of -ITIS words [10]. It draws in part on the same data sources but includes suffixes not treated in [9, 10]. The methods are similar though not identical. With respect to the most recent publication of the NIH group [9], our results corroborate theirs where the same suffixes are treated. In addition, the present work differs in its specific application of morphosemantic analysis to the creation of dictionary entries for automatic text processing.

Linguistic Analysis

1. Sublanguage Vocabulary

Sublanguages differ from common English both with respect to their syntax [4, 6, 13] and their lexicon. The difference is not one of essence but one of degree. Just as certain syntactic patterns commonly found in general English might be absent in the sublanguage texts while others might have a comparatively higher or lower frequency of occurrence, there are words that are part and parcel of general English that rarely or never occur in these texts, and vice versa.

The vocabulary of clinical narrative can be subdivided into three general classes:

I General English vocabulary

Lexical items of this group include articles, prepositions, conjunctions, quantifiers and words or affixes expressing time or negation. They generally do not carry sublanguage-specific information and are therefore not given medical subclasses in the LSP system.

II English vocabulary with sublanguage-specific usage

Here belong words of the major lexical classes that, although part of everyday vocabulary, are used in well-defined and much narrower senses in the sublanguage. The noun *foot,* for instance, which in common English denotes a bodypart, the bottom (of a mountain, of a page, etc), a unit of length, and a unit of metrical verse, among others, denotes exclusively a bodypart in medical texts.

III Medical terminology

This group comprises all words specific to the subject area. Although items from this group might find their way into popular language—witness the use of terms like *appendectomy, oncology, amniocentesis* by laymen—this portion of the vocabulary of clinical reporting is generally reserved for the exchange of information among experts in the field. A prerequisite for its understanding and proper usage is extensive knowledge of the subject area. Words of this group are almost exclusively nouns and adjectives.

Words of the first class do not exhibit formal regularities suitable for automatic classification. As members of relatively small and closed word classes they have largely been coded for the LSP dictionary.

Lexical items of the second group can be given their English and sublanguage subclasses upon careful investigation of their cooccurrence restrictions in the documents. This is particularly true for verbs, which, almost without exception, are general English verbs used in a well-defined sense in the medical documents. Manual classification is of course necessary for all monomorphemic items such as *jaw, show, kidney, tender,* but is also required with regard to morphologically complex items, e.g., *treatment, enlargement, examination, irregularity* and others. Although the latter can be assigned to the class of singular nouns and the subclass of nonhuman nouns because of the suffixes *-ment, -ion,* and *-ity* that regularly form abstract singular nouns from verbal or adjectival bases, these suffixes are unrevealing as to the medical subclass membership of the complex nouns containing them. Compare the entries of just four forms containing the suffix *-ment* as listed in the medical dictionary of the LSP*.

(NSIX)	treatment
.11 =	NONHUMAN, H-VTR: (GENERIC)
(NSIX)	improvement
.11 =	NONHUMAN, H-CHANGE, H-STATUS
(NSIX)	enlargement
.11 =	NONHUMAN, H-CHANGE, H-INDIC
(NSIX)	appointment
.11 =	NONHUMAN, H-VMD

While many morphologically complex words share the medical subclass with their bases—both *treat* and *treatment* are H-VTR words, to take but one example—, the relationship between the medical subclass of the base and the derived word is not systematic.

The vocabulary of the third class differs markedly from standard English by the prevalence of words of Greek and Latin extraction. Many of these strictly medical terms have come down to us from antiquity unchanged; they have always been part of the medical vocabulary. But due to the terminological needs of the field many words or word parts that had non-medical referents in the classical languages have been endowed with medical meanings, either by metaphoric extension or simply by convention. As often as not there is a discrepancy between the etymological meaning of a word or word part and its present-day technical meaning. Lexical items that once were independent words have acquired the status of affixes, e.g., Greek *raphe* 'seam' is now considered a suffix by medical dictionaries. Affixes such as *-asis* or *-osis*, which in classical Greek carried as much (or as little) information as their English equivalents *-ion* or *-ment* have come to carry quite specific semantic information in medical terminology. Thus *-asis* and *-osis* words consistently denote an abnormal condition, *X-oma* words almost invariably refer to a tumor or growth of some sort; *-itis*, originally an adjective suffix equivalent to our *-ic* or *-al* has developed into a noun-forming suffix denoting an inflammation. This portion of the vocabulary abounds furthermore in hybrids (e.g. *retinoscopy*, or *dysgerminoma*) and linguistic improprieties such as the use of the nominative rather than oblique stem of Greek words in combinations (*iritis* rather than *iriditis*). Clearly, words belonging to this group can no longer be measured by the standards of the classical languages. Knowledge of classical Greek or Latin, though useful, therefore does not greatly facilitate their dictionary classification.

Simplex words of this third class, e.g. *salpinx* or *coccyx*, have to be classified manually just like native simplex items. Neo-Latin complex words, on the other hand, are eminently suited for machine analysis. This is so because they are not only morphologically transparent, i.e. analyzable into their constituent parts, but also semantically transparent. *Encephalitis* and *adenoma*, for instance, can be segmented into the constituents *encephal* and *itis*, and *aden(o)* and *oma*, with their meaning being a function of the meaning of their parts. More important, perfect compositionality is not limited to newly coined medical terms but is also present in established *(lexicalized)* terms. In contradistinction, English complex words are perfectly compositional only as long as they are novel. With usage they acquire additional shades of meaning or altogether new meanings so that there is no longer a one-to-one mapping between the constituents and their referents. Consequently, only the classes and subclasses of novel complex words of class II can be predicted with accuracy, not however those of established words. The category and subclasses of both lexicalized and novel Neo-Latin medical terms, on the other hand, can be inferred from their constituents.

2. Medical Morphology

Whereas complex words of general English are primarily the result of concatenating two free forms *(houseboat, tapeworm)* or a free and a bound form *(preexisting, redness)*, medical

Table 1 Definitions of the Medical Subclasses Mentioned in this Text

Subclass	Definition
H-AREA	word is generally associated with bodypart words and identifies an area of a bodypart *(vertex, tip)*
H-BODYPART	word denotes a bodypart or organ. The class includes body fluids and secretions *(liver, peritoneum, urine)*
H-CHANGE	word indicates a change in an existing state *(improvement, increase, progress)*
H-DIAG	word denotes a disease. The ICDA manual has been used as the basis for classification *(arthritis, German measles)*
H-GROW	word relates to the patient's growth or development *(grow, birth)*
H-INDIC	item is a member of the class of disease-indicator words *(fever, swelling)*
H-LOC	item specifies the location of an area of a bodypart *(lower, bilateral)*
H-ORG	item denotes an organism grown or cultured out in the course of a lab test *(cocci, pathogen, microorganism)*
H-PROC	item denotes a test procedure which is performed on the patient in contrast to lab tests *(shunt-o-gram, electrocardiogram)*
H-STATUS	word reflecting current status of patient's condition with respect to disease in question *(convalesce, stable, exacerbate)*
H-SURG	item refers to a therapeutic surgical procedure *(neurosurgery, vasectomy)*
H-TIMELOC	item marks relative position in time *(prior, postpartum, prenatal)*
H-VMD	word of general medical management, relates the medical institution or personnel to the patient, the diagnosis, or treatment of the patient *(admission, examine)*
H-VTR	word denotes a therapeutic action (as opposed to a medication, which is H-RX) that is prescribed or given by the medical staff *(treatment, bed rest)*

* The LSP system replaces recurrent references to the same type of category list with a short canonical form, e.g. (NSIX) identifies the word to the right as a Noun Singular with additional subclasses on a prespecified numbered line (e.g. .11 for noun subclasses). Similarly, (NPLX) stands for a plural noun with subclasses, (ADJX) for an adjective with subclasses. The LSP medical subclasses referred to in this paper are defined in Table 1 below. An early version of the LSP medical subclasses was given in [5].

terminology consists, to a large extent, of words that are composed of two bound forms *(gastritis, cardiomegaly)*. This type of concatenation is dealt with in the literature on word-formation rather cursorily under the heading *Neo-Latin compounding* [8] or root-compounding. It is given short shrift because it calls into question the established separation of word-formation processes into derivation (bound form + free form) and composition (free form + free form).

The distinction between affixes and Neo-Latin combining forms and between derivation and composition is of considerable theoretical interest to linguists but can be disregarded for our purposes. Note that no matter how we label the individual components of complex words, it is always the last constituent that determines their major lexical category. Thus *treatment, boathouse,* or *underdog* are nouns because their last constituent is a noun. Similarly, *arteriosclerosis, rhinoplasty,* and *dermatitis* are nouns because their last (or *head*) constituent assigns the category noun, independent of whether it is classified as a free form, a combining form, or a bona fide suffix.

The head constituent determines in addition the subclasses relevant for information processing. This is quite obviously the case when the last constituent is itself a free form, as in *arteriosclerosis* which has the same English and medical subclasses *as sclerosis*. Compare their LSP dictionary entries below (see previous footnote for abbreviations):

(NSIX)	sclerosis
.11 =	NONHUMAN, H-DIAG
(NSIX)	arteriosclerosis
.11 =	NONHUMAN, H-DIAG

Thus given the entry *sclerosis,* the dictionary worker only has to transfer the lexical class and subclass information to a word analyzable as *X-sclerosis*. (With regard to the additional information provided by the *X*-portion of the word, see below.)

This is also true for complex words whose last constituent has no free counterpart. The consistency with which *mastectomy, vasectomy, appendectomy,* and, in fact, any word analyzable as *Xectomy* are specified as nonhuman count nouns (NCOUNT1) belonging to the class of surgery words (H-SURG) allows us to give *Xectomy* words the general dictionary schema

(NSIX)	Xectomy
.11 =	NONHUMAN, NCOUNT1, H-SURG

Given this schema, a dictionary worker need not be an expert in the field to correctly classify any noun conforming to this pattern.

The word segment *-ectomy* does not represent an isolated instance of a recurrent terminal sequence conveying very specific lexical and subclass information. Table 2 provides a partial listing of terminal constituents and their associated dictionary schemata that can be used by dictionary workers as a template for classifying complex words exhibiting those patterns.

Beyond regularities of the type above, regularities of a much subtler nature can be found in complex medical words. Thus, words with a given terminal sequence that yield the expected entries have consistently first constituents of a particular semantic subclass. For instance, the constituent preceding the terminal sequence *-tomy* refers in practically all cases to an object which has the medical subclass H-BODYPART in the LSP system. Similarly, the sequence preceding *-rrhea* refers to a bodyfluid (*pyorrhea* 'pus') or a bodypart (*gastrorrhea* 'stomach'), both of which are subclassified as H-BODYPART for the purposes of the LSP analyzer. In many cases the initial constituent is itself complex. *Meningoencephalopathy* is composed of the elements *meningo-* and *encephalo-*, both of which designate bodyparts, plus *-pathy;* in *meningoencephalomyopathy* three elements denoting bodyparts precede the terminal sequence.

All of the terminal sequences identified in Table 2 allow their coconstituent(s) to be of the semantic subclass H-BODYPART. Other semantic types may be involved, as for *-osis* in Table 3. However, in the corpus surveyed (see Section 3 below) there is no instance of a terminal sequence that cooccurs with first constituents of all theoretically possible subclasses. Table 3 illustrates that, apart from morphological structure, medical terms have a rather clearcut semantic structure (cf. also [9, 10]).

As indicated earlier, the first constituent of a complex word is not decisive for its dictionary classification, since the relevant properties are determined by its last constituent. Access to semantic information supplied by first constituents is nevertheless valuable for retrieval of information below the word level, hence should not be ignored. More important, the identification of stable semantic cooccurrence patterns and the ranking of competing patterns permits us to handle words of a specified terminal sequence whose dictionary classifications do not conform to the schemata established in Table 2. *Anatomy* and *autotomy,* for instance, do not denote surgical procedures. They deviate from the predominant semantic pattern of *-tomy* words in that their first constituents are not of the semantic type H-BODYPART. Pacak et al. [10] cite *mephitis* as an instance of an *-itis* word that does not denote an inflammation. This is because *mephitis* diverges from the pattern of *-itis* words that require a constituent denoting a bodypart to precede the terminal constituent.

3. Methods of Obtaining Affix Data

For some of the patterns in Table 2 we drew on the findings of Pacak et al. [10]. Additional patterns were established based on inspection of a sample corpus: when a number of words with a recurrent terminal sequence held the prospect of pattern regularity, these forms were checked for semantic coherence and an attempt was made to find more examples to test the predictive reliability of the hypothesized pattern. For this phase, a

Table 2 Terminal constituents and associated dictionary entry schemata

Suffix		Entry Schema
-rrhea	(NSIX) .11 =	Xrrhea NONHUMAN, H-DIAG
-rrhagia	(NSIX) .11 =	Xrrhagia NONHUMAN, H-DIAG
-oma	(NSIX) .11 =	Xoma NONHUMAN, H-DIAG
-cele	(NSIX) .11 =	Xcele NONHUMAN, H-DIAG
-gram	(NSIX) .11 =	Xgram NONHUMAN, H-PROC
-ptosis	(NSIX) .11 =	Xptosis NONHUMAN, H-DIAG
-sepsis	(NSIX) .11 =	Xsepsis NONHUMAN, H-DIAG
-phraxis	(NSIX) .11 =	Xphraxis NONHUMAN, H-DIAG
-rrhexis	(NSIX) .11 =	Xrrhexis NONHUMAN, H-DIAG
-oncus	(NSIX) .11 =	Xoncus NONHUMAN, H-DIAG
-rrhaphy	(NSIX) .11 =	Xrrhaphy NONHUMAN, H-SURG
-plasty	(NSIX) .11 =	Xplasty NONHUMAN, H-SURG
-pexy	(NSIX) .11 =	Xpexy NONHUMAN, H-SURG
-itis	(NSIX) .11 =	Xitis NONHUMAN, H-DIAG
-algia	(NSIX) .11 =	Xalgia NONHUMAN, H-INDIC
-emia	(NSIX) .11 =	Xemia NONHUMAN, H-DIAG
-osis	(NSIX) .11 =	Xosis NONHUMAN, H-DIAG
-scopy	(NSIX) .11 =	Xscopy NONHUMAN, H-PROC
-otomy	(NSIX) .11 =	Xotomy NONHUMAN, H-SURG
-ectomy	(NSIX) .11 =	Xectomy NONHUMAN, H-SURG
-ostomy	(NSIX) .11 =	Xostomy NONHUMAN, H-SURG
-centesis	(NSIX) .11 =	Xcentesis NONHUMAN, H-PROC
-megaly	(NSIX) .11 =	Xmegaly NONHUMAN, H-INDIC

(Plural forms have been omitted)

SPITBOL program that scans the boldface entries in Dorland's Medical Dictionary [1] and extracts all items ending in a specified terminal sequence proved very useful, since it enabled us to test the hypothesis against a sizable body of data. The output of the SPITBOL program provided at the same time valuable statistics in that it gave us a general idea of the number of forms exhibiting a given pattern. Thus Dorland's lists more than 2000 words ending in *-itis,* close to a thousand words in *-tomy,* several hundred in *-emia,* to name just the largest groups.

We also employed a computational method of pattern establishment. Rather than extracting words with like terminal constituents that were specified beforehand, we ran a program that extracts and groups words that are identical with respect to the last n letters. Groups were then tested for semantic coherence and transparency as before.

Program Design

The morphology program described below translates the strategies employed in the manual analysis and classification of Neo-Latin medical terms into procedures that are executable by the computer. It reads an input word and determines whether the word's terminal character string belongs to the set of word segments carrying lexical category and subclass information. The output is a fully

Table 3 Illustrative Sample of Semantic Patterns

Initial constituents					Last constituent		
other	H-BODY-PART	H-TES-TYPE	H-DIAG	H-BODY-PART	H-BODY-PART		
					angi-o	-oma	H-DIAG
				angi-o	lip	-oma	H-DIAG
					cardi-o	-gram	H-PROC
electro					cardi-o	-gram	H-PROC
					mening-o	-pathy	H-DIAG
				mening-o	encephal-o	-pathy	H-DIAG
	mening-o			encephal-o	my-o	-pathy	H-DIAG
					aden-	-osis	H-DIAG
		alkal				-osis	H-DIAG
	angi		-oma-t-			-osis	H-DIAG
					myx-o	-rrhea	H-DIAG
					ur-o	-rrhea	H-DIAG
				colp-o	episi-o	-rrhaphy	H-SURG
					celi-o	-rrhaphy	H-SURG
					arthr-o	-plasty	H-SURG
					enter-o	-plasty	H-SURG
					hepat-o	-ptosis	H-DIAG
					cec-o	-ptosis	H-DIAG
					nephro	-pexy	H-SURG
				my-o	nephr-o	-pexy	H-SURG
					blephar	-oncus	H-DIAG
par					ophthalm	-oncus	H-DIAG
					hemat	-oma	H-DIAG
				hem	angi	-oma	H-DIAG
				pancreat-o	duoden-o	-stomy	H-SURG
				ureter-o	pyel-o	-stomy	H-SURG
				abdomin-o	hyster-o	-tomy	H-SURG
					blephar-o	-tomy	H-SURG
					tonsill	-ectomy	H-SURG
				pylor-o	gastr	-ectomy	H-SURG
					pyel-o	-plasty	H-SURG
					rhin-o	-plasty	H-SURG
					thorac-o	-centesis	H-PROC
					pleur-o	-centesis	H-PROC
					rect	-algia	H-INDIC
					peritone	-algia	H-INDIC
					appendic	-itis	H-DIAG
litho					nephr	-itis	H-DIAG

coded entry suitable for updating the existing medical dictionary of the LSP. Unanalyzable words are returned for manual processing.

The first step toward this goal is a "suffix" dictionary that associates terminal constituents with their canonical form and their English and medical subclasses. Using the principle of the longest match, an input word is scanned from right to left for any of the sequences listed in the suffix dictionary. If a match is achieved, the lexical information supplied by the suffix is transferred to the input word.

The successful match between the last n letters of an input word and a suffix is sufficient in many cases to obtain correct updates. However, for *atremia, asemia, coma,* or *achroma* it would produce incorrect results, for here a matched character string does not represent a terminal constituent. Misanalyses of this kind can be ruled out only if the word is exhaustively analyzed into all its constituents.

Formal clues do exist. The connecting vowel *-o-* (occasionally *-i-*) frequently signals a constituent break (e.g. *nephro + pexy, resini + ferous*). It cannot be depended upon in a strictly formal procedure because (i) *-o-*'s occur in places other than the edge of constituents, (ii) the connecting vowel is omitted when the following constituent begins with a vowel or certain consonants *(ostealgia, oophorrhagia).*

A solution to this problem is to supplement the suffix dictionary with a dictionary of first constituents or "prefixes". Once a terminal character string has been identified, a successful match between the remainder of the word and an item on the prefix list guarantees that the word is analyzed into its proper constituents rather than into a suffix and an arbitrary and meaningless sequence of characters.

Dorland's Medical Dictionary lists the most frequently occurring bound forms that are found as first constituents of Neo-Latin terms. Entries for the prefix dictionary can be extracted from this source with the aid of the SPITBOL program mentioned earlier, searching for all items ending in "-" in this case. To ensure successful matches for e.g., *nephritis (nephr-), nephropexy (nephro-), dermolysis (dermo-),* prefixes are listed in all their variant forms. As Neo-Latin bound forms often differ quite drastically from their free counterparts—compare *nephr-* and *kidney,* or *rhin-* and *nose*—and as they carry quite specific semantic information that we might want to access, prefixes are listed with their associated medical subclass and their free counterparts in the prefix dictionary.

One also has to take into account that the sequence preceding the suffix may itself be complex. Polymorphemic constituents such as *pericardi-* that refer to well-defined entities are included in the prefix dictionary. In *nephropyeloplasty* the suffix is preceded by *nephro-* and *pyelo-,* in *nephroureterostomy* by *nephro-* and *ureter-,* in *nephrouretercystectomy* by *nephro-, ureter-* and *cyst-.* Listing *nephropyelo-, nephroureter-* and *nephrouretercyst-* as items on the prefix list would result in an unduly cumbersome prefix list.

These considerations led us to adopt the following scanning strategy: The suffix routine is invoked only once since properties relevant to dictionary creation are signalled by the suffix alone. For a complete constituent analysis, the remainder of the word is scanned from left to right for a match with an item from the prefix dictionary. The prefix routine is invoked repeatedly so as to permit the analysis of words with multiple prefixes (e.g., *cholangiohepatitis, nephroureterostomy*). It is called independently of the suffix routine to allow for partial analyses of words even if no suffix match can be achieved. Although partial analyses of this kind will not result in updates, they provide valuable clues as to which suffixes and prefixes have to be incorporated into the experimental affix dictionary.

The advantage of a bidirectional over a unidirectional scan is that it takes the two definite affix boundaries into account, viz. the first letter of the prefix and the last letter of the suffix. It also reduces the likelihood of incomplete analyses or misanalyses inherent in unidirectional scans using the principle of the longest match. Pacak et al. [10] cite the case of *panotitis,* segmented by their right to left analysis as *pa-not-itis,* i.e. into an unmatched sequence *pa* and the matched sequences *not* and *itis.* Assuming that *pan- ot-, not-,* and *-itis* are listed in the affix dictionary, a bidirectional scan will correctly analyze the word into *pan-ot-itis.*

Provisions had to be made to rule out, or at least reduce, misanalyses due to juncture phenomena. The word *oophorrhagia,* for example, has the constituents *oophor* and *rrhagia,* but would be analyzed into **oopho* and *rrhagia.* To allow for an overlap of graphemes of this kind, the remainder of the word is to be redefined after a successful affix match to include the first (or last) letter of the matched affix. (cf. also [10]). This ensures that a complex word whose constituents are listed in the affix dictionaries receives an exhaustive analysis on the second pass.

First pass	*oopho + rrhagia
Second pass	oophor + rrhagia
First pass	phlebo + *phthalmo + *tomy
Second pass	phlebo + ophthalmo + otomy

More extensive overlaps, as in *pancreatomy* from *pancreato-* and *-otomy* have not been taken into account but have been considered exceptional.

To curtail misanalyses of the type *a-r-oma* or *re-ct-algia,* which are expected to be prevalent due to incompleteness of the pilot affix dictionary, we associate with each affix a value indicating the number of characters the remainder of the word has to contain for the affix match to be considered successful. The default value is set at 2, but can be overridden when necessary. Setting the value to 5 for *a-,* for example, rules out the prefix *a-* in *aroma.* This is by no means a cure-all, but it aborts at least some attempted matches that are bound to fail so long as the affix dictionary is incomplete.

A very specific decision guiding the design of the suffix dictionary was to list suffixes in their longest recurrent

form. For instance, given the suffix entry *-tomy* and the prefix entries *nephr-/nephro-, thoraco-/thorac-* and the input words *nephrotomy, nephrectomy,* and *thoracostomy,* only *nephrotomy* would receive a complete analysis. Listing *-ostomy* and *-ectomy,* which are complex suffixes, and *-otomy* with the combining vowel *-o-* as part of the suffix, allows all these words to be analyzed exhaustively, or leaves us in the case of a partial analysis with a remainder that is more likely a prefix.

A modular program written in FORTRAN 77, and PASCAL, consisting of an affix (or rules) dictionary, a rule compiler, an affix search controller and an LSP specific processor creating the desired output files has been implemented on a Digital Equipment Corporation Vax 11/780 running VMS. Adding affixes to the rules dictionary is a straightforward procedure which requires no prior knowledge of any programming language. The program can be run interactively or batch. In the first case, the user will define a word using the terminal for input and output. In this mode, the program is a look-up tool that enables the user to obtain grammatical and semantic information about an input word. When used batch, the user submits input words in the form of a file. The program returns (1) a file of successfully analyzed words with their dictionary classifications that can be used directly to update the existing medical dictionary of the LSP; (2) a file of unanalyzed words that have to be coded manually; and (3) a file tracing the analysis of every input word into its constituents and meaning components. A diagnostic file of this type enables the user to check on the correctness of the analyses and aids in determining what additional affixes have to be incorporated into the affix dictionary.

Tests and Results

An automatic dictionary classification program is a viable alternative to manual classification only if it processes sufficiently many words and yields reliable entries. We tested the reliability of the program with a small rules dictionary containing 10 suffixes and approximately 200 prefixes. A substantially larger rules dictionary was used in one of the coverage tests.

1. Reliability Tests

First we extracted from the LSP medical dictionary all words ending in the suffixes specified by the pilot program and compared the manually prepared entries with the machine-generated ones. The comparison showed that a substantial portion of the test sample (74%) was classified as well, or even better, by the program than by hand. It also pointed out omissions and inconsistencies of manual processing and suggested ways for improving the program.

For a second reliability check we ran the program against a large list of words ending in a specified terminal sequence and checked the output manually. The purpose of this experiment was to establish whether correct updates crucially depend on an exhaustive analysis of an input word into its constituent parts or whether a successful suffix match alone would yield sufficiently dependable results. We extracted from Dorland's sections a-b and p-v all words ending in *-tomy,* as this would enable us to test the suffix routines for *-otomy, -ostomy,* and *-ectomy.* The test results are listed in Table 4.

348 words, or 87% of the sample words tested, were given correct entries by the program. The partially analyzed words indicated that the experimental rules dictionary had to be substantially augmented. They also showed that correct updates—at least of *-ostomy, -otomy,* and *-ectomy* words—were not necessarily contingent on their exhaustive analysis. 26 words were not analyzed because they contained more than 20 characters. This technical restriction, in effect at the time, does not pertain to the adequacy of the program.

Two factors were found to be responsible for the incorrect analyses:

1) The word does not conform to the prevalent semantic pattern of *-tomy* words which require a H-BODY-PART constituent to precede the suffix. This is the case for *autotomy,* whose prefix has no sublanguage-specific subclass, and also for *bdellotomy* and *blastotomy,* whose first constituents denote an object other than a bodypart.
2) The word has a first constituent that influences the subclass assignment: *postmastectomy, postcardiotomy* are time expressions, referring to a period after surgery rather than to a surgical procedure.

Both problems could be handled by a more sophisticated program that incorporates restrictions to test for the semantic compatibility of co-constituents.

Table 4 Test of Dorland's sample for the suffixes *(-otomy, -ostomy,* and *-ectomy)*

Total number of Dorland's words tested	399	100%
Number of correctly classified words	348	87%
Number of exhaustively analyzed words	119	30%
Number of partially analyzed words yielding correct updates	229	57%
Number of incorrectly classified words	12	3%
Number of words rejected by program	39	10%
too long	26	
no suffix match: e.g. *anatomy*	13	

2. Coverage Tests

With respect to coverage we determined the proportion of words analyzable by machine in relation to the total number of words that had to be classified in order to process a new set of documents. We found that on the average 12% to 14% of these words could be coded automatically with the small experimental program. This figure represents approximately two-thirds of the Neo-Latin words in the »not-found« group. The remaining (non-Neo-Latinate) portion consisted largely of abbreviations and eponyms.

The program, with a substantially enlarged rules dictionary containing 25 suffix and 778 prefix definitions, was run against an arbitrarily selected portion of Dorland's Medical Dictionary (entries *abbe* to *aedoeocephalus*).

The words of the test sample are grouped by morphological types in Table 5.

Table 5 Morphological types in Dorland's sample (entries *abbe* through *aedoeocephalus*)

Total number of words		480
Duplicates		160
Total number of distinct words		320
Total number of non-analyzable forms		44
Abbreviations	14	
Prefixes	6	
Eponyms	10	
Simplex words	14	
Total number of complex words		276
with productive English morphology	78	
with Neo-Latin morphology	198	
nouns	134	
adjectives	64	

In Table 5 the term "duplicates" refers to entries of the type *adenoma fibrosum, adenoma psammosum* or *fetal adenoma,* which are subentries of the flush-left boldface entry *adenoma* in the printed version of Dorland's. Subentries consisting of a word preceded or followed by a qualifying word are treated as idioms by LSP conventions. Their lexical category, English and medical subclasses are identical to that of the head word. Once we subtract these unanalyzable simplex words and abbreviations from the total number of words, we are left with a rather large number of complex words. A characteristic of the technical vocabulary of medicine is the predominance of Neo-Latin complex words which outweigh complex words built according to productive patterns of general English at a ratio of 2:1.

In this test the program processed 56 (42%) of the 134 Neo-Latin nouns it was intended to handle. This poor showing deserves further comment. A persistent problem, even with our enlarged rules dictionary, was its incompleteness. The addition of only 3 suffix entries would have brought up the number of analyzed words to 69 (51%). Neo-Latin complex adjectives cannot be processed by the program in its present form. This is because the analysis of Neo-Latin complex adjectives requires more refined strategies than the analysis of nouns. A more complete affix dictionary, in conjunction with the procedure for coding adjectives that is currently being implemented will considerably increase the percentage of words that can be coded by machine.

Conclusions

Preliminary linguistic analysis showed that for a large number of Neo-Latin medical terms LSP-type dictionary entries of considerable detail can be inferred from such formal lexical properties as the presence of specified terminal sequences. The identification of these terminal constituents and the establishment of dictionary schemata as exemplified in Table 2 make it possible to classify such medical terms into broad informational categories without reference to their narrative context, and, in fact, without a specialist's knowledge of the meaning of these terms. These features make Neo-Latin medical terms excellent candidates for machine classification.

Our experiments with a pilot dictionary classification program confirmed the supposition that reliable dictionary entries reflecting the meaning of components of Neo-Latin complex forms could be generated by machine. Coverage and reliability of the program can be increased by incorporating a restriction component that tests coconstituents for their semantic compatibility. This will enable us to eliminate, in a principled manner, those words that do not conform to the semantic schemata as illustrated in Table 3. The addition of procedures for classifying Neo-Latin complex adjectives will permit us to capture the large number of medical adjectives. Yet, no matter how small the percentage of misclassified words may be, the output of the dictionary classification program must still be checked manually before updating the text processing dictionary to ensure that no incorrect entries are added.

The program is useful for the small dictionary updates that have to precede the processing of new documents as well as for more massive updates that take, for instance, all of the words listed in Dorland's Medical Dictionary as input.

In view of the openendedness of the medical vocabulary, the program should prove useful for the analysis and classification of medical terms that are not yet defined in the standard dictionaries. It can therefore help one to keep abreast of the terminological innovations in the field.

Because of the international character of the Neo-Latin vocabulary of medicine, adapting the program for a German or French medical dictionary, for instance, would involve little more than altering the spelling of some affixes.

While these efforts do not eliminate manual classification altogether, they are invaluable in reducing one major obstacle to the large-scale application of natural language processing to medical narrative.

Acknowledgements

This research was supported in part by National Library of Medicine grant number 1-RO1-LM03933, awarded by the National Institutes of Health, Department of Health and Human Services.

We are grateful to the W. B. Saunders Co. for making available a computer-readable list of the words in Dorland's Illustrated Medical Dictionary, 25th edition. Carol Foster is responsible for the computer implementation.

REFERENCES

[1] Dorland's Illustrated Medical Dictionary: 25th edition., (Philadelphia: W. B. Saunders 1974).

[2] Fitzpatrick, E., Sager, N.: The Lexical Subclasses of the Linguistic String Parser. American Journal of Computational Linguistics, microfiche 2 (1974). Reprinted in: String Program Reports 9, New York University Linguistic String pp. 322–374 (Reading, Mass.: Addison-Wesley 1981).

[3] Friedman, C., Sager, N., Chi. E., Marsh, E., Christenson, C., Lyman, M.: Computer Structuring of Free Text Patient Data. Proceedings of the 7th Annual Symposium on Computer Applications in Medical Care (SCAMC7). Washington, D.C., October 1983.

[4] Harris, Z. S.: Mathematical Structures of Language. (New York: Wiley Interscience 1968).

[5] Hirschman, L., Sager, N.: Automatic Information Formatting of a Medical Sublanguage. In Kittredge, R., Lehrberger, J. (Eds): Sublanguage: Studies of Language in Restricted Semantic Domains, pp. 27–80. (Berlin: de Gruyter 1982).

[6] Kittredge, R.: Variation and Homogeneity of Sublanguages. In Kittredge, R., Lehrberger, J. (Eds): Sublanguage: Studies of Language in Restricted Semantic Domains, pp. 108–137. (Berlin: de Gruyter 1982).

[7] Longman Dictionary of Contemporary English. (London: Longman Ltd. 1978).

[8] Marchand, H.: The Categories and Types of Present-Day English Word-Formation. (München: C. H. Becksche Verlagsbuchhandlung 1969).

[9] Norton, L. M., Pacak, M. G.: Morphosemantic Analysis of Compound Word Forms Denoting Surgical Procedures. Meth. Inform. Med. *22* (1983) 29–36.

[10] Pacak, M. G., Norton, L. M., Dunham, G. S.: Morphosemantic Analysis of -ITIS Forms in Medical Language. Meth. Inform. Med. *19* (1980) 99–105.

[11] Pacak, M. G., Pratt, A. W.: Identification and Transformation of Terminal Morphemes in Medical English, Part II. Meth. Inform. Med. *17* (1978) 95–100.

[12] Pratt, A. W., Pacak, M. G.: Identification and Transformation of Terminal Morphemes in Medical English. Meth. Inform. Med. *8* (1969) 84–90.

[13] Sager, N.: Syntactic Formatting of Science Information. AFIPS Conference Proceedings 41. pp. 791–800 (Montvale, NJ: AFIPS Press 1972) Reprinted in Kittredge, R., Lehrberger, J. (Eds): Sublanguage: Studies of Language in Restricted Semantic Domains. pp. 9–26. (Berlin: de Gruyter 1982).

[14] Sager, N.: Natural Language Information Processing: A Computer Grammar of English and its Applications. (Reading, Mass.: Addison-Wesley 1981).

[15] Sager, N.: Natural Language Information Formatting: The Automatic Conversion of Texts to a Structured Data Base. In Yovits, M. C. (Ed.): Advances in Computers Vol. *17*, pp. 89–162 (New York: Academic Press 1978).

[16] Stedman's Medical Dictionary, 21st edit. (Baltimore: Williams and Wilkins Co. 1966).

[17] Webster's Seventh Collegiate Dictionary. (Springfield, Mass.: G. & C. Merriam Company 1967).

[18] Webster's Third New International Dictionary. (Springfield, Mass.: G. & C. Merriam Company 1971).

[19] White, C., Foster, C.: Automatic Parsing with a Hybrid Lexicon. In String Program Reports 13, New York University Linguistic String Project (1980).

[20] White, C.: The Linguistic String Project Dictionary for Automatic Text Analysis. Proceedings of the Workshop on Machine Readable Dictionaries (Menlo Park: SRI International 1983).

Address of the author:
Susanne Wolff, Ph. D., New York University,
Linguistic String Project,
251 Mercer Street
New York, N. Y. 10012, USA