UMLS® Reference Manual

UMLS® Reference Manual

SPECIALIST Lexicon and Lexical Tools

[6]

The SPECIALIST Lexicon has been developed to provide the lexical information needed for the SPECIALIST Natural Language Processing System (NLP). It is intended to be a general English lexicon that includes many biomedical terms. Coverage includes both commonly occurring English words and biomedical vocabulary. The lexicon entry for each word or term records the syntactic, morphological, and orthographic information needed by the SPECIALIST NLP System.

The Lexical Tools are designed to address the high degree of variability in natural language words and terms. Words often have several inflected forms which would properly be considered instances of the same word. The verb "treat", for example, has three inflectional variants: "treats" the third person singular present tense form, "treated" the past and past participle form, and "treating" the present participle form. Multi-word terms in the Metathesaurus and other controlled vocabularies may have word order variants in addition to their inflectional and alphabetic case variants. The Lexical Tools allow the user to abstract away from this sort of variation.

For an overview of the SPECIALIST Lexicon, lexical variant programs, and lexical databases, see Lexical Methods for Managing variation in Biomedical Terminologies, A.T. McCray, S. Srinivasan, A.C. Browne, in the Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994, 235-239.

The SPECIALIST Lexicon is distributed as one of the UMLS Knowledge Sources and as an open source resource along with the the SPECIALIST NLP tools, subject to these terms and conditions.

6.1 General Description

The Lexicon consists of a set of lexical entries with one entry for each spelling or set of spelling variants in a particular part of speech. Lexical items may be "multi-word" terms made up of other words if the multi-word term is determined to be a lexical item by its presence as a term in general English or medical dictionaries, or in medical thesauri such as MeSH. Expansions of generally used acronyms and abbreviations are also allowed as multi-word terms.

The unit lexical record is a frame structure consisting of slots and fillers. Each lexical record has a base= slot whose filler indicates the base form, and optionally a set of spelling_variants= slots to indicate spelling variants. An "entry=" slot records the unique identifier (EUI) of the record. EUI numbers are seven digit numbers preceded by an "E". Each record has a cat= slot indicating part of speech. The lexical record is delimited by braces ({...}).

The unit lexical records for "anaesthetic" given below illustrate some of the features of the SPECIALIST lexical record:

{base=anesthetic spelling_variant=anaesthetic entry=E0330018 cat=noun variants=reg variants=uncount } {base=anesthetic spelling_variant=anaesthetic entry=E0330019 cat=adj variants=inv position=attrib(3) position=pred stative }

The base form "anesthetic" and its spelling variant "anaesthetic" appear in two lexical records, one an adjective entry, the other a noun entry. The variants= slot contains a code indicating

the inflectional morphology of the entry; the filler reg in the noun entry indicates that the noun "anesthetic" is a count noun which undergoes regular English plural formation ("anaesthetics"); inv in the variants= slot of the adjective entry indicates that the adjective "anesthetic" does not form a comparative or superlative. The position= slot indicates that the adjective "anaesthetic" is attributive and appears after color adjectives in the normal adjective order. "pred" in the position slot of the adjective entry indicates that this adjective can appear in predicate position.

Lexical entries are not divided into senses. Therefore, an entry represents a spelling-category pairing regardless of semantics. The noun "act" has two senses both which show a capitalized and lower case spelling; an act of a play and an act of law. Since both senses share the same spellings and syntactic category, they are represented by a single lexical entry in the current lexicon. The unit record for "Act" is shown below.

{base=Act spelling_variant=act entry=E0000154 cat=noun variants=reg }

When different senses have different syntactic behavior, codes for each behavior are recorded in a single entry. For example, "beer" has two senses: the alcoholic beverage and the amount of a standard container of that beverage.

A. Patients who drank beer recovered more slowly than patients who drank wine.
B. Fifty-six patients reported drinking more than five beers a day.

The first sense illustrated in A. above is a mass (uncount) noun. The second sense illustrated in B. is a regular (count) noun. In cases like this the appropriate codes for both senses are included in the entry.

{base=beer entry=E0012226 cat=noun variants=uncount variants=reg }

Two codes will also appear in cases where the lexical item is both count and uncount without a sense distinction. "Abdominal delivery" denotes the same procedure whether it appears as an uncount noun as in C. or a count noun as in D.

C. Abdominal delivery is the procedure of choice in this situation.
D. Abdominal deliveries are more common these days.

The unit lexical record for "abdominal delivery" includes both codes.

{base=abdominal delivery entry=E0006453 cat=noun variants=uncount variants=reg }

Other syntactic codes such as complement codes for verbs, adjectives and nouns are similarly grouped without regard to sense.

6.2 The Scope of the Lexicon

Words are selected for lexical coding from a variety of sources. Approximately 20,000 words from the UMLS Test Collection of MEDLINE abstracts together with words which appear both in the UMLS Metathesaurus and Dorland's Illustrated Medical Dictionary form the core of the words entered. In addition, an effort has been made to include words from the general English vocabulary. The 10,000 most frequent words listed in The American Heritage Word Frequency Book and the list of 2,000 words used in definitions in Longman's Dictionary of Contemporary English have also been coded. Since the majority of the words selected for coding are nouns, an effort has been made to include verbs and adjectives by identifying verbs in current MEDLINE citation records, by using the Computer Usable Oxford Advanced Learner's Dictionary, and by identifying potential adjectives from Dorland's Illustrated Medical Dictionary using heuristics developed by McCray and Srinivasan (1990).

A variety of reference sources are used in coding lexical records. Coding is based on actual usage in the UMLS Test Collection and MEDLINE, dictionaries of general English, primarily learner's dictionaries which record the kind of syntactic information needed for NLP, and medical dictionaries. Longman's Dictionary of Contemporary English, Dorland's Illustrated Medical Dictionary, Collins COBUILD Dictionary, The Oxford Advanced Learner's Dictionary, and Webster's Medical Desk Dictionary were used.

The SPECIALIST Lexicon also exists in relational format generated from the unit records. The full SPECIALIST Lexicon technical report entitled "The SPECIALIST Lexicon," found in the file techrpt.pdf, fully describes the unit record format. The remainder of the present chapter describes the relational form of the Lexicon. Section 6.3 describes the data elements that make up the relational tables and Section 6.4 describes the tables.

6.3 Lexicon Data Elements

Each of the elements below is represented as fields (columns) in the relational format.

6.3.1 String Properties

These data elements refer to properties of the strings generated by the entries.

6.3.1.1 STR - String

A Lexical entry generates a variety of forms (strings) including all the inflectional forms (the citation form, as well) of each spelling variant. Case, punctuation and spaces are considered significant.

6.3.1.2 AGR - Agreement/Inflection Code

This element encodes agreement and inflection information.

Agreement between nouns and verbs and between determiners and nouns involves person and number. Person and Number are indicated by the following codes.

Code	Person	Number
second	Second	Singular & Plural
third	Third	Singular & Plural
fst_sing	First	Singular
fst_plur	First	Plural
thr_sing	Third	Singular
thr_plur	Third	Plural

For Nouns, the agreement/inflection code indicates countability, person and number. Person and number are indicated by the person/number codes given above which are parenthesized after the countability code. Nouns can be either count or uncount.

For Pronouns, the agreement/inflection indicates person and number using the codes given above.

For verbs, including auxiliaries and modals, the agreement/inflection code indicates tense, person and number. Persons and numbers are indicated by the same person/number codes given above. These codes are parenthesized after the tense. No person number codes are given for non-finite tenses. "pres(thr_sing)" indicates third person singular present tense and "pres (fst_sing,fst_plur,thr_plur,second)" indicates present tense for all persons and numbers other

than third singular. Negative forms of auxiliaries (didn't) and modals (can't) have "negative" after a colon at the end of the agreement/inflection code.

Code	Tense
past	Past Tense
pres	Present Tense
past_part	Past Participle
pres_part	Present Participle
infinitive	Infinitive

Determiners agree with nouns in terms of countability and number. The agreement/inflection codes for determiners are "free", "plur", "sing" and "uncount". "free" indicates that the determiner places no restrictions on its noun. Determiners marked "plur" allow plural nouns, those marked "sing" allow singular nouns and those marked "uncount" allow uncountnouns.

6.3.1.3 CAS - Case

See Section 4.3.1 of "The SPECIALIST Lexicon" technical report.

Pronouns in English may be in one of two cases, subjective (nominative) or objective (accusative). This field contains "subj", "obj" or both separated by a comma to indicate the case of the pronoun.

6.3.1.4 GND - Gender

This field indicates the gender of pronouns.

Pronouns may be marked pers or neut to indicate whether they refer to people or non-people respectively. Pronouns marked pers may be masculine (masc) or feminine (fem) referring to male or female people respectively. See Section 14.2 of "The SPECIALIST Lexicon" technical report. There are four codes possible in this field:

Code	Gender
pers	person
neut	neuter
pers(masc)	person masculine
pers(fem)	person feminine

Notice that pers as used here does not correspond to the traditional term "personal pronoun". For example "it" and "they" are traditionally called personal pronouns since they both participate in the person/number paradigm. A pronoun like "none" is not traditionally called a personal pronoun.

6.3.2 Entry Properties

6.3.2.1 EUI - Unique Identifier Number for Lexical Entries

The EUI identifies a lexical entry. Information about a set of spelling variants in a particular part of speech is represented as an entry in the unit record. A particular string may be assigned several EUI numbers as it may occur in several parts of speech.

6.3.2.2 CIT - Citation Form

This field records the citation form of strings in the agreement/inflection table (Section 6.4.3.1 - lragr). The citation form is the singular for nouns, infinitive for verb and positive for adjectives and adverbs. The base form and the spelling variants if any are the citation forms of each of their respective inflections. This form is sometimes referred to as the un-inflected form.

6.3.2.3 BAS - Base Form

This field records the base form of a lexical entry. The base form is the citation form of one of a set of spelling variants chosen to represent the whole set. It might be thought of as the name of a lexical entry. The base form is the filler of the base= slot.

6.3.2.4 SCA - Syntactic Category

The syntactic category (part of speech) of the lexical entry. This field may be filled by one of the following. See Section 3 of "The SPECIALIST Lexicon" technical report.

Code	Category
noun	nouns
adj	adjectives
adv	adverbs
pron	pronouns
verb	verbs
det	determiners
prep	prepositions
conj	conjunctions
aux	auxiliaries
modal	modals
compl	complementizers

6.3.2.5 PER - Periphrastic

The code "periph" in this field indicates that an adjective or adverb is periphrastic. An adjective is periphrastic if it can form its comparative with "more" and its superlative with "most". See Section 4.3.5 of "The SPECIALIST Lexicon" technical report for discussion.

6.3.2.6 COM - Complements

These are complement codes. See Sections 5.1, 5.2, 5.4 and 5.5 in "The SPECIALIST Lexicon" technical report for a description of SPECIALIST complement codes.

6.3.2.7 TYP - Inflectional Type

The inflectional type(s) of an entry indicate the ways in which its forms may be inflected, or in the case of determiners the inflection of the heads they may determine. These codes are used to generate the variant strings (STR) found in other tables.

For nouns the following types may appear:

Code	Pluralization Pattern	See "The SPECIALIST Lexicon" Section
reg	regular	4.5.2
glreg	Greco-Latin regular	4.5.3
metareg	metalinguistic regular	4.5.4
irreg( )	irregular	4.5.5
sing	fixed singular	4.5.6
plur	fixed plural	4.5.7
inv	invariant	4.5.8
group(irreg( ))	group irregular	4.5.9
group(reg)	group regular	4.5.9
uncount	uncountable	4.5.10
groupuncount	group uncount	4.5.11

For verbs the following types may appear:

Code	Inflection Type	See "The SPECIALIST Lexicon" Section
reg	regular	4.1.1
regd	regular doubling	4.1.2
irreg( )	irregular	4.1.3

For pronouns the following types may appear:

Code	Inflection Type
fst_plur	first person plural
fst_sing	first person singular
sec_plur	second person plural
sec_sing	second person singular
second	second person
third	third person
thr_plur	third person plural
thr_sing	third person singular

See Section 14.1 of "The SPECIALIST Lexicon" technical report.

For adjectives and adverbs the following types can appear:

UMLS® Reference Manual

UMLS® Reference Manual

Code	Inflectional Type	See "The SPECIALIST Lexicon" Section
reg	regular	4.3.1 and 4.4.1
regd	regular doubling	4.3.2
inv	invariant	4.3.4 and 4.4.3
inv;periph	periphrastic	4.3.5 and 4.4.4
irreg( )	irregular	4.3.3 and 4.4.2

For determiners the inflection type indicates the inflection of the noun heads they may determine. The following types may appear:

Code	Inflectional Type	See "The SPECIALIST Lexicon" Section
sing	singular	4.7.1
plur	plural	4.7.2
uncount	uncount	4.7.3
singuncount	singular uncount	4.7.4
pluruncount	plural uncount	4.7.5
free	free	4.7.6

6.3.2.8 POS - Possession

English pronouns may be possessive or possessive nominal. The codes poss, possnom or both (comma separated) may appear in this field.

See Section 14.3.2 of "The SPECIALIST Lexicon" technical report.

6.3.2.9 QNT - Quantification

This field indicates the quantification properties inherent in certain pronouns. The four codes possible in this field are:

Code	Properties
univ	universal quantification
indef(nonassert)	non-assertive indefinite
indef(neg)	negative indefinite
indef(assert)	assertive indefinite

See Section 14.3.4 in "The SPECIALIST Lexicon" technical report for discussion of quantification in pronouns.

6.3.2.10 FEA - Features

This field represents various features of terms in various categories. The possible features are:

Feature	See "The SPECIALIST Lexicon" Section
reflexive	14.3.3
negative	14.3.4
demonstrative	14.3.5
interrogative	12.1
proper	8.
negative	13.1
broad_negative	13.2
stative	10.

6.3.2.11 PSN - Position for Adjectives

Adjectives are marked in the SPECIALIST Lexicon with position codes showing whether they are attributive postmodifying or predicative. If attributive, the code indicates where they appear in the pre-nominal sequence of adjectives. An additional attributive code, attribc, is used to indicate adjectives which can take complements in attributive position. One or more of the following codes can appear:

Code	Position	See "The SPECIALIST Lexicon" Section
attrib(1)	attributive (1st position)	9.1.1.1
attrib(2)	attributive (2nd position)	9.1.1.2
attrib(3)	attributive (3rd position)	9.1.1.3
attribc	attributive with complement	9.1.2
post	post modifying	9.2
pred	predicative	9.3

6.3.2.12 MOD - Modification Type for Adverbs

Adverbs are marked in the SPECIALIST Lexicon to indicate their modification type. The possible values of this field are:

Code	See "The SPECIALIST Lexicon" Section
intensifier	11.2
particle	11.1
sentence_modifier; TYPE	11.3
verb_modifier; TYPE	11.4

TYPE is one of locative, temporal or manner. See Section 11.5 in "The SPECIALIST Lexicon" technical report.

6.3.2.13 GEN - Generic Name for a Trademark

The GEN field represents a generic or public name for the thing referred to by the trademark. The trademark "Alphalin" has the generic term "vitamin A".

6.3.3 Entry Relations

6.3.3.1 ABR - Acronym or Abbreviation

This field indicates whether a term listed in the acronym-abbreviation table (lrabr) is an acronym or abbreviation. It contains either:

"abbreviation_of" or "acronym_of".

6.3.3.2 SPV - Spelling Variant

A base form in the SPECIALIST Lexicon may have one or more spelling variants, subject to the same inflectional pattern. This field contains the citation form of a particular spelling variant. See Section 2 of "The SPECIALIST Lexicon" technical report.

6.3.4 Data Description

The data elements describe the relational table files or provide index entries into the Lexicon.

6.3.4.1 WRD - Word

Each string is broken into "words" and indexed in lrwd. Words are strings of alpha-numeric characters more than one character long, separated by space or punctuation.

6.3.4.2 DES - Description

A short definition of a file or field. This is free text.

6.3.4.3 FMT - Format

An ordered comma separated list of field names appearing in a file.

6.3.4.4 RWS - Number of Rows

The number of rows (lines or records) in a file.

6.3.4.5 FIL - File Name(s)

One or more file names denoting the files containing relational tables.

6.3.4.6 BTS - Size in Bytes

The size of a file in bytes (characters).

6.3.4.7 CLS - Number of Columns

The number of columns (fields) in a record (or row) of a table. The same number as the number of lines in the file.

6.3.4.8 COL - Three Letter Field Name

A three letter identifier for a field.

References

1.

6.3.4.9 REF - Cross Reference to Document

A cross reference to a section of this document.

6.4 Lexicon Relational Tables

6.4.1 Introduction

In this format the data in each lexical entry is represented in ten different "relations" or "tables" each in a file.

The lexicon relational format is not fully normalized. By design, there is duplication of data among different relations and within certain relations. Developers will need to make their own decisions about the extent to which this redundancy should be retained, reduced, or increased for their specific applications.

6.4.2 General Description of the Relational Format

As in the Metathesaurus ASCII relational format, each relation or table of data values has by definition a fixed number of columns; the number of rows depends on the content of a particular version of the Lexicon. A column is a sequence of all the values in a given data element or logical sub-element. In general, columns for longer variable length data elements will appear to the right of columns for shorter and/or fixed length data elements. A row contains the values for one or more data elements or logical sub-elements for one Lexicon entry or string. Depending on the nature of the data elements involved, each Lexicon entry or string may have one or more rows in a given file. The values for the different data elements or logical sub-elements represented in the row are separated by vertical bars (|). If an optional element is blank, the vertical bars are still used to maintain the correct positioning of the subsequent elements. Each row is terminated by a vertical bar and a carriage return followed by a line feed. (|<CR><LF>).

6.4.3 Summary of the Contents of Each of the Relational Files

In the following descriptions, the numbers in parentheses beside each element refer to the section of this document that describes the element's contents.

6.4.3.1 - Agreement and Inflection (File = lragr)

Rows of the agreement table have six fields. There is a row in lragr for each inflected form of each spelling variant. This table links those forms to their citation forms and base forms. It provides information about agreement between subjects (nouns and pronouns) and verbs and between determiners and nouns.

EUI	The Entry Unique ID Number (6.3.2.1)
STR	String (6.3.1.1)
SCA	Syntactic Category (6.3.2.4)
AGR	Agreement/Inflection Code (6.3.1.2)
CIT	Citation Form (6.3.2.2)
BAS	Base Form (6.3.2.3)

6.4.3.2 - Inflection Type (File = lrtyp)

The lrtyp table has one or more rows for each lexical entry, indicating the inflectional pattern (s) to which it belongs.

UMLS® Reference Manual

UMLS® Reference Manual

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
SCA	Syntactic Category (6.3.2.4)
TYP	Inflectional Type (6.3.2.7)

6.4.3.3 - Complementation (File = lrcmp)

In lrcmp there is one line for each complement code for each entry.

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
SCA	Syntactic Category (6.3.2.4)
COM	Complement Code. (6.3.2.6)

6.4.3.4 - Pronouns (File = lrprn)

lrprn has one or more rows for each pronoun entry in the Lexicon. Each row has nine columns.

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
AGR	Agreement/Inflection Code (6.3.1.2)

See Section 14.1 in "The SPECIALIST Lexicon" technical report.

The agreement/inflection field in lrprn indicates person and number for anaphoric reference, AGR in lragr indicates person for agreement. These differ in the case of possessive nominal pronouns. The possessive nominal "mine" is "third" for purposes of subject verb agreement and "fst_sing" in its anaphoric reference.

GND	Gender (6.3.1.4)
CAS	Case (6.3.1.3)
POS	Possession (6.3.2.8)
QNT	Quantification (6.3.2.9)
FEA	Other Features (for pronouns) (6.3.2.10)

6.4.3.5 Modifiers (file = lrmod)

The modifier table includes position information for adjectives and modification type information for adverbs, and a variety of features.

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
SCA	Syntactic Category (6.3.2.4)

All the entries represented in this table have the category "adj" or "adv" indicating adjectives or adverbs respectively.

PSN/MOD

The fourth field of lrmod may be one of the following depending on whether the term is an adjective or adverb.

PSN	Position (6.3.2.11) - for adjectives
MOD	Modification Types (6.3.2.12) - for adverbs
FEA	Features (6.3.2.10)

6.4.3.6 - Properties (file = lrprp)

lrprp indicates properties of terms in various categories.

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
SCA	Syntactic Category (6.3.2.4)
STR	String (6.3.1.1)

STR is only indicated in lrprp when a feature applies to a single string out of those generated by the entry, as in the negative contractions.

FEA	Features (6.3.2.10)

6.4.3.7 - Abbreviations and Acronyms (file = lrabr)

This file links acronyms and abbreviations to their expansions.

EUI	The Entry Unique ID Number (6.3.2.1)

This field contains the EUI of the acronym or abbreviation.

BAS	The Base Form (6.3.2.3)

This field contains the Base form of the acronym or abbreviation.

ABR	Acronym or Abbreviation (6.3.3.1)
BAS	The Base Form (6.3.2.3)

This field contains the Base form of the expansion of the acronym or abbreviation.

EUI	The Entry Unique ID Number (6.3.2.1)

This field contains the EUI of the expansion of the abbreviation or acronym.

6.4.3.8 - Spelling Variants (file = lrspl)

EUI	The Entry Unique ID Number (6.3.2.1)
SPV	Spelling Variant (6.3.3.2)
BAS	The Base Form (6.3.2.3)

6.4.3.9 - Nominalizations (file = lrnom)

This file contains the EUI of the nominalization.

BAS	The Base Form (6.3.2.3)

This field contains the Base form of the nominalization.

SCA	Syntactic Category (6.3.2.4)

This field contains the category of the nominalization (noun).

EUI	The Entry Unique ID Number (6.3.2.1)

This field contains the EUI of a verb or adjective of which the noun is a nominalization.

BAS	The Base Form (6.3.2.3)

This field contains the base form of the verb or adjective of which the noun is a nominalization.

SCA	Syntactic Category (6.3.2.4)

This field contains the syntactic category (adj or verb) of the adjective or verb.

6.4.3.10 - Trademarks (file = lrtrm)

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	Base (6.3.2.3)
GEN	Generic Term (6.3.2.13)

The appearance of a form in the lrtrm table indicates that it is a trademark. It may or may not have a generic term associated with it.

6.4.3.11 - Files (file = lrfil)

The lrfil table describes each file in the ASCII relational form of the Lexicon.

FIL	File Name(s) (6.3.4.5)
DES	Description (6.3.4.2)
FMT	Format (6.3.4.3)
CLS	Number of Columns (6.3.4.7)
RWS	Number of Rows (6.3.4.4)
BTS	Size in Bytes (6.3.4.6)

6.4.3.12 - Word Index. (file = lrwrd)

WRD	Word (6.3.4.1)
EUI	The Entry Unique ID Number (6.3.2.1)

UMLS® Reference Manual

UMLS® Reference Manual

6.4.3.13 - Fields (file = lrfld)

COL	Three Letter Field Name (6.3.4.8)
DES	Description (6.3.4.2)
REF	Cross Reference to Document (6.3.4.9)
FIL	File Name(s) (6.3.4.5)

6.5 The SPECIALIST Lexicon Unit Record

The unit lexical record is a frame structure consisting of slots and fillers. Each lexical record has a base= slot whose filler indicates the base form, and optionally a set of spelling_variants= slots to indicate spelling variants. Lexical entries are delimited by entry= slots filled by the EUI number of the entry. EUI numbers are seven digit numbers preceded by an "E". Each entry has a cat= slot indicating part of speech. The lexical record is delimited by braces ({...}).

The unit lexical records for "anaesthetic" given below illustrate some of the features of a SPECIALIST unit lexical record:

{base=anesthetic spelling_variant=anaesthetic entry=E0330018 cat=noun variants=reg variants=uncount } {base=anesthetic spelling_variant=anaesthetic entry=E0330019 cat=adj variants=inv position=attrib(3) position=pred stative }

The base form "anesthetic" and its spelling variant "anaesthetic" appear in two lexical records containing a noun and a verb entry. The variants= slot contains a code indicating the inflectional morphology of the entry; the filler reg in the noun entry indicates that the noun "anaesthetic" is a count noun which undergoes regular English plural formation ("anaesthetics"); inv in the variants= slot of the adjective entry indicates that the adjective "anesthetic" does not form a comparative or superlative. The position= slot indicates that the adjective "anaesthetic" is attributive and appears after color adjectives in the normal adjective order.

The SPECIALIST technical report "The SPECIALIST Lexicon" gives a full description of the Lexicon in unit format.

6.6 Lexical Databases Introduction

The lexical databases contain lexical information that we have found to be useful for Natural Language Processing. They are not finished products but are under continuous development.

6.6.1 Semantically Related Terms SM.DB

This database (SM.DB) contains pairs of semantically related terms. Each row of the database has the following form.

TERM1|SCA1|TERM2|SCA2

Such a row indicates that TERM1 in syntactic category SCA1 is semantically related to TERM2 in syntactic category SCA2. Both terms are given in base form.

Examples:

alar|adj|wing|noun

ocular|adj|eye|noun

auditory area|noun|auditory cortex|noun

vomitive|noun|emetic|noun

vomitive|adj|emetic|adj

iridescent virus|noun|iridovirus|noun

typhloteritis|noun|cecitis|noun

6.6.2 Derivationally Related Terms: DM.DB

This database (DM.DB) contains pairs of terms related by derivational morphology. Each row of the database has the same form as sm.db. Both terms are given in base form.

TERM1|SCA1|TERM2|SCA2

Examples:

abashment|noun|abash|verb
adenohypophyseal|adj|adenohypophysis|noun
amenorrheic|adj|amenorrhea|noun
arithmetician|noun|arithmetic|noun
convert|verb|conversible|adj
immobilize|verb|immobility|noun

DM.DB is derived from the morphological fact files (dm.fct, etc.) used in lvg (See Lexical Variant Generation section in Section 6.8).

6.6.3 Spelling Variants: SP.DB

The Spelling Variant database (SP.DB) contains pairs of terms that are spelling variants of each other. The format of each row is the same as the format of dm.db and sm.db. SCA1 and SCA2 are always the same in SP.DB.

TERM1|SCA1|TERM2|SCA2

Examples:

accouter|verb|accoutre|verb
accurst|adj|accursed|adj
acidaemic|adj|acidemic|adj
aesthetics|noun|esthetics|noun
dairy farmer|noun|dairy-farmer|noun

SP.DB is derived from the SPECIALIST Lexicon.

6.6.4 Neo-classical Combining Forms NC.DB

This database (NC.DB) contains morphemes that are used to form neo-classical compounds. Each row of the database has the following form.

MORPHEME|MEANING|TYPE

Morphemes may have optional connecting vowels indicated in parentheses. The types are: prefix, root, and terminal.

Examples:

abdomin(o)|abdomen|root
ab|away from|prefix
acou(o)|hearing|root
cardi(o)|heart|root
cele|swelling|terminal
desis|binding|terminal
de|negate|prefix

Our analysis of combining forms divides them into roots and terminals, which are distinguished from prefixes and suffixes. A neo-classical compound can consist of any number of roots ending in a terminal or suffix. Prefixes normally must precede roots and cannot attach directly to terminals. Users interested in suffixation rules and facts should consult the dm.rul and dm.fct files included with lvg.

For further discussion see McCray et. al., 1988, "The Semantic Structure of Neo-Classical Compounds", In the Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care, Washington DC.

6.7 Sample Records

::::::::::::::
lragr.sam
::::::::::::::
E0007127|acute|adj|positive;periph|acute|acute|
E0014875|cans|noun|count(thr_plur)|can|can|
E0014875|can|noun|count(thr_sing)|can|can|
E0014876|canned|verb|past_part|can|can|
E0014876|canned|verb|past|can|can|
E0014876|canning|verb|pres_part|can|can|
E0014876|cans|verb|pres(thr_sing)|can|can|
E0014876|can|verb|infinitive|can|can|
E0014876|can|verb|pres(fst_sing,fst_plur,thr_plur,second)|can|can|
E0014877|can't|modal|pres:negative|can|can|
E0014877|cannot|modal|pres:negative|can|can|
E0014877|can|modal|pres|can|can|
E0014877|couldn't|modal|past:negative|can|can|
E0014877|could|modal|past|can|can|
E0014937|canine teeth|noun|count(thr_plur)|canine tooth|canine tooth|
E0014937|canine tooth|noun|count(thr_sing)|canine tooth|canine tooth|
E0017902|colors|noun|count(thr_plur)|color|color|
E0017902|color|noun|count(thr_sing)|color|color|
E0017902|color|noun|uncount(thr_sing)|color|color|
E0017903|colored|verb|past_part|color|color|
E0017903|colored|verb|past|color|color|
E0017903|coloring|verb|pres_part|color|color|
E0017903|colors|verb|pres(thr_sing)|color|color|
E0017903|color|verb|infinitive|color|color|
E0017903|color|verb|pres(fst_sing,fst_plur,thr_plur,second)|color|color|
E0051632|quickly|adv|positive;periph|quickly|quickly|
E0055585|she|pron|thr_sing|she|she|

::::::::::::::
lrcmp.sam

UMLS® Reference Manual

UMLS® Reference Manual

::::::::::::::
::::::::::::::
E0014876|can|verb|tran=np|
E0017903|color|verb|cplxtran=np,adj|
E0017903|color|verb|cplxtran=np,np|
E0017903|color|verb|intran;part(in)|
E0017903|color|verb|intran;part(up)|
E0017903|color|verb|intran|
E0017903|color|verb|tran=np;part(in)|
E0017903|color|verb|tran=np|

::::::::::::::
lrmod.sam
::::::::::::::
E0007127|acute|adj|attrib(1),attrib(3),pred|stative|
E0051632|quickly|adv|verb_modifier;manner||

::::::::::::::
lrnom.sam
::::::::::::::
E0007121|acuity|noun|E0007127|acute|adj|
E0021126|deduction|noun|E0021123|deduce|verb|
E0021126|deduction|noun|E0021124|deduct|verb|
E0061851|transportation|noun|E0061850|transport|verb|

::::::::::::::
lrprn.sam
::::::::::::::
E0030918|he|thr_sing|pers(masc)|subj||||
E0036100|it|thr_sing|neut|subj,obj||||
E0055585|she|thr_sing|pers(fem)|subj||||

::::::::::::::
lrprp.sam
::::::::::::::
E0007127|acute|acute|adj|stative|
E0004825|Parkinson|Parkinson|noun|proper|
E0014877|can|can't|modal|negative|
E0014877|can|can't|modal|negative|
E0014877|can|couldn't|modal|negative|

::::::::::::::
lrspl.sam
::::::::::::::
E0017902|colour|color|
E0017903|colour|color|
E0008769|anesthetic|anaesthetic|
E0008770|anesthetic|anaesthetic|

::::::::::::::
lrtrm.sam
::::::::::::::
E0412633|Actinex|meso-nordihydroguaiaretic acid|
E0415286|Antivert||

UMLS® Reference Manual

UMLS® Reference Manual

E0414928|thioplex|thiotepa|
E0415019|theo-hexanicit||

::::::::::::::
lrtyp.sam
::::::::::::::
E0007127|acute|adj|inv;periph|
E0014875|can|noun|reg|
E0014876|can|verb|regd|
E0014937|canine tooth|noun|irreg(canine teeth)|
E0017902|color|noun|reg|
E0017902|color|noun|uncount|
E0017903|color|verb|reg|
E0051632|quickly|adv|inv;periph|

::::::::::::::
lrwd.sam
::::::::::::::
acute|E0001203
acute|E0007127
acute|E0007130
acute|E0007131
acute|E0007132
acute|E0007133
acute|E0007134
acute|E0007135
acute|E0007136
acute|E0007137
acute|E0007138
acute|E0007139
acute|E0007140
acute|E0007141
acute|E0007142
acute|E0007143
acute|E0007144
acute|E0007145
acute|E0007146
acute|E0007147
acute|E0007148
acute|E0007149
acute|E0007150
acute|E0007151
acute|E0007152
acute|E0007153
acute|E0007154
acute|E0007155
acute|E0007156
acute|E0007157
acute|E0007158
acute|E0007159
acute|E0007160
acute|E0007161
acute|E0007162

acute|E0007163
acute|E0007164
acute|E0007165
acute|E0007166
acute|E0007167
acute|E0007168
acute|E0007169
acute|E0007170
acute|E0007171
acute|E0007172
acute|E0007173
acute|E0007174
acute|E0007175
acute|E0007176
acute|E0007177
acute|E0007178
acute|E0007179
acute|E0007180
acute|E0007181
acute|E0007182
acute|E0007183
acute|E0007184
acute|E0007185
acute|E0007186
acute|E0007187
acute|E0007188
acute|E0007189
acute|E0007190
acute|E0007191
acute|E0007192
acute|E0007193
acute|E0007194
acute|E0007195
acute|E0007196
acute|E0007197
acute|E0007198
acute|E0007199
acute|E0007200
acute|E0007201
acute|E0007202
acute|E0007203
acute|E0007204
acute|E0007205
acute|E0007206
acute|E0007207
acute|E0007208
acute|E0007209
acute|E0007210
acute|E0007211
acute|E0007212
acute|E0007213
acute|E0007214
acute|E0007215

UMLS® Reference Manual

UMLS® Reference Manual

acute|E0016430
acute|E0018044
acute|E0019256
acute|E0200089
acute|E0200090
acute|E0203254
acute|E0208423
acute|E0208433
acute|E0208452
acute|E0208475
acute|E0208494
acute|E0210443
acute|E0210574
acute|E0210575
acute|E0210576
acute|E0210642
acute|E0214476
acute|E0216615
acute|E0216616
acute|E0217176
acute|E0217376
acute|E0217551
acute|E0217756
acute|E0313307
acute|E0314926
acute|E0319558
acute|E0321232
acute|E0321304
acute|E0322005
acute|E0332592
acute|E0409630
acute|E0418090
acute|E0418484
acute|E0418485
acute|E0418705
acute|E0420121
acute|E0422597
acute|E0422634
acute|E0422824
acute|E0422825
can|E0014875
can|E0014876
can|E0014877
can|E0014875
can|E0014876
can|E0014877

6.8 The SPECIALIST Lexical Tools

The SPECIALIST Lexical Tools package consists of three primary programs -- a normalizer, a word index generator, and a lexical variant generator, together with a set of ancillary programs for normalization. This package is implemented in Java.

The SPECIALIST Lexical Tools and the SPECIALIST Lexicon are distributed as one of the UMLS Knowledge Sources and along with the SPECIALIST NLP Tools as open source resources subject to these terms and conditions.

Updates and bug fixes can be found in the release notes on the Download Lexical Tools Web page.

The distributions come with install programs (for Solaris, Linux, and Window) and a ReadMe.txt file describing how to install and configure the Lexical Tools and providing a brief description of each program.

The **docs** directory contains user guides, Java API documents, and design documents describing in detail the use of Lexical Tools. This document is a general introduction to the programs in the lexical variant generation package.

The compressed Lexical Tools are as follows*:

lvg2008.tgz

- The official distribution of lvg. This includes the source code for the programs, the data and tables in a pure Java embedded database (Instant DB) the programs use, full documentation, installation instructions, and jar files of the programs. See the documents contained within this distribution for a more complete description of this product.

*File names for the 2008 release are shown.

Normalization (norm)

The lexical program **norm** generates the normalized strings that are used in the normalized string index, MRXNS. Thus norm must be used before MRXNS can be searched.

The normalization process involves stripping possessives, replacing punctuation with spaces, removing stop words, lower-casing each word, breaking a string into its constituent words, and sorting the words in alphabetic order. The uninflected forms are generated using the SPECIALIST Lexicon if words appear in the Lexicon, otherwise they are generated algorithmically. When a form could be an inflection of more than one base form, the new normalization process returns multiple uninflected forms. If a string to be normalized contains multiple ambiguous forms, and the permutation of these ambiguous forms offer more than 10 output forms, the input form lowercased, with punctuation replaced, word order sorted, but not uninflected, is returned. The upper limit of permutation number (10) is configurable by modifying the configuration file. The program **luiNorm** has the behavior of prior year's normalization, and is distributed for those who need it.

Norm reads its standard input and writes to standard output. It expects input lines to be records separated into fields. The field separator is |. The string to be normalized is identified to norm using the **-t** option. **-t** takes a numerical argument which denotes the field in which the input string is to be found. If no **-t** option appears, norm assumes that the input string is in the first field (**-t:1**). There need not be more than one field, so lines consisting only of input strings are properly understood.

Norm output records include all the fields of the input record with an additional field to the right containing the normalized form of the input string.

For example, if the user had a list of terms to be looked up via the normalized string index in a file called **terms**, he or she could use **norm -i:terms -o:terms.nrm** to get the normalized form of each term. If the input file **terms** contained the following:

2, 4-Dichlorophenoxyacetic acid
Syndrome, anterior, compartment
Abnormal, weight, gain
Anemia, Refractory, with Excess of Blasts
left atriums

the file **term.nrm** would contain:

2, 4-Dichlorophenoxyacetic acid|2 4 acid dichlorophenoxyacetic
Syndrome, anterior, compartment|anterior compartment syndrome
Abnormal, weight, gain|abnormal gain weight
Anemia, Refractory, with Excess of Blasts|anemia blast excess refractory
left atriums|atrium left
left atriums|atrium leave

The string in the second field of each line of terms.nrm is now suitable for matching to MRXNS.

Word Index (wordInd)

The lexical program **wordInd** breaks strings into words for use with the word index in MRXW. Users of the word index should use wordInd to break strings into words before searching in the word index. This assures congruence between the words to be looked up and the word index.

Word for this purpose is defined as a token containing only alphanumeric characters with length one or greater. The wordInd program lowercases the output words.

The wordInd program reads its standard input and writes to its standard output. Like norm and lvg, it expects each input line to be a record separated into fields by |. The field containing the input string is identified using the **-t** option. The numerical argument of **-t** denotes the field in which the input string may be found. If no **-t** option is given, the input string is expected to be in the first field (**-t:1**). There need not be more than one field, so lines consisting only of input strings are properly understood.

The wordInd program outputs one line of output for each word found in the input string. Input fields are not repeated in the output unless specified in a **-F** option. Applying wordInd to the input string **Heart Disease, Acute** would result in three output lines:

heart
disease
acute

The numerical argument of **-F** indicates an input field to be repeated in the output. A numerical argument for **-F** option is required for each input field that is to be repeated. Fields are repeated in the order in which the numerical argument of **-F** options appear. The output words always appear as an additional field to the right of any repeated input fields. For example, applying **wordInd -t:2 -F:2:1** to a record of the form **UI23456|tooth, canine|definition.....**; would result in the following output:

tooth, canine|UI23456|tooth
tooth, canine|UI23456|canine

The third field of each of those records contains a word extracted from the input term in the first field (**-t:2** ,**-F:2**). The **-F:1** option repeats the UI numbers from the first field of input. The fact that **-F:2:1** placed the UI numbers (field 1) after the input string (field 2).

Lexical Variant Generation (lvg)

The lvg program generates lexical variants of input words. It consists of several different flow components that can be combined in various ways to produce lexical variants. The user of lvg chooses combinations of flow components and combines them into a **flow**. (The normalizer program, norm, is essentially the lvg program with a pre-selected flow option: **lvg -f:N**.) The arguments of the **-f** flag are used to specify a flow. Each flow can be thought of as a pipeline with each flow component feeding the next. For example, the flow **-f:i** simply generates inflectional variants and **-f:l:i** generates lowercase inflectional variants. Each of the flow components options is discussed on the documents for lvg.

The lvg program reads from its standard input and writes to its standard output. Input records may be typed in at the keyboard, after typing the command on the command line (**lvg -f:i**) or input lines may be read from a file (**lvg -f:i -i:file**) or piped to lvg from another command (**COMMAND|lvg -f:i**). Output records may be directed to the screen (default), send to a file (**lvg -f:i -i:INFILE -o:OUTFILE**) or piped to another command (**lvg -f:i -i:infile | COMMAND**).

Input

The lvg program is designed to work with one line input records divided into fields. The default field separator is |. The field separator can be changed using the **-s** option. The field in which the input term, whose variants are to be generated, can be specified with the **-t** option. In the absence of a **-t** flag the input term is assumed to be in the first field of the input. So both **dog** and **dog|canine|UI4567** would generate variants of **dog**. With the **-t** flag set to **2**, **dog|canine|UI4567** would generate variants of **canine**. In the case of single field input (**dog**), lvg generates variants from the only field regardless of the setting of **-t**.

The lvg program can read category (part of speech) and inflection information from the input record. The numerical argument to the **-cf** option indicates the field in which category information is located. In the input record, category information needs to be encoded as a number according to the scheme described on the documents for lvg. The numerical argument to the **-if** option indicates the field in which inflection information is located. In the input record, inflection information needs to be encoded as a number according to the scheme described on the documents for Lexical Tools.

Output

The lvg program adds five new fields to the input record and outputs a record for each variant generated. For example, if **dog|canine|UI4567** is given to the standard input of **lvg -f:i** the output sent to standard out will be:

dog|canine|UI4567|dog|128|1|i|1| dog|canine|UI4567|dog|128|512|i|1| dog|canine|UI4567|dogs|128|8|i|1| dog|canine|UI4567|dog|1024|1|i|1| dog|canine|UI4567|dog|1024|262144|i|1| dog|canine|UI4567|dog|1024|1024|i|1| dog|canine|UI4567|dogs|1024|128|i|1| dog|canine|UI4567|dogged|1024|64|i|1| dog|canine|UI4567|dogged|1024|32|i|1| dog|canine|UI4567|dogging|1024|16|i|1|

The first three fields of each record above are identical to the input record, the rest are supplied by lvg. The first additional field is the variant form lvg has generated. The second additional

field is the syntactic category of the variant encoded as a number. The third additional field is the inflection of the variant encoded as a number. The fourth additional field indicates the flow that was selected. The fifth field is the number of the flow which generated this variant. Output category (parts of speech) and inflection information are encoded in the same scheme used for input category and inflection information.

Further description of the SPECIALIST Lexical Tools is available at the SPECIALIST Lexical Tools Web site: http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html.