The SPECIALIST Lexicon

The SPECIALIST lexicon is one of three UMLS® Knowledge Sources under development by the National Library of Medicine (NLM) as part of the Unified Medical Language System® project.

The SPECIALIST lexicon has been developed to provide the lexical information needed for the SPECIALIST Natural Language Processing System. 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 natural language processing system.

Scope and Content of the SPECIALIST lexicon

Lexical entries may be single or multi-word terms. Each lexical record has a base form a part of speech, a unique identifier and optionally a set of spelling variants. The base form is the uninflected form of the lexical item; the singular form in the case of a noun, the infinitive form in the case of a verb, and the positive form in the case of an adjective or adverb.

Lexical information includes syntactic category, inflectional variation (e.g., singular and plural for nouns, the conjugations of verbs, the positive, comparative, and superlative for adjectives and adverbs), and allowable complementation patterns (i.e., the objects and other arguments that verbs, nouns, and adjectives can take). The lexicon recognizes eleven syntactic categories, or parts of speech: verbs, nouns, adjectives, adverbs, auxiliaries, modals, pronouns, prepositions, conjunctions, complementizers, and determiners.

The basic sentence patterns of a language are determined by the number and nature of the complements taken by verbs. The lexicon recognizes five broad complementation patterns: intransitive, transitive, ditransitive, linking and complex-transitive. Verb entries also encode each of the inflected forms (principal parts of the verb). Verbs are inflectionally classified as regular, Greco-Latin regular or irregular. Noun entries describe the inflection of the nouns (pluralization) and spelling variations. Complementation patterns for nouns and nominalization information are also included when relevant. In addition to inflection and complement codes, adjectives in the lexicon have position codes to indicate the syntactic positions in which they may occur. An adjective may be a qualitative, classifying, or color adjective. Adverbs in the lexicon are coded to indicate their modification properties. The lexicon recognizes sentence, verb phrase and intensifier type adverbs, and classifies sentence and verb phrase adverbs into manner, temporal and locative types.

Lexical items are selected for coding from a variety of sources, including lexical items from MEDLINE® citation records, and a large set of lexical items from medical and general English dictionaries.

Distribution Formats

The SPECIALIST lexicon is provided in two formats; a unit record format and a relational table format.

The information associated with each lexical entry includes a unique identifier, a base form, a syntactic category code, certain agreement information, complementation information if relevant, and various other properties relevant to the particular lexical entry.

The unit record format is a frame structure consisting of slots and fillers. The slots are the basic lexical attributes, and the fillers express the possible values of those attributes for that particular lexical item.

Data for lexical entries are also represented in a set of relational tables. 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 decide the extent to which this redundancy should be retained, reduced, or increased for their applications. Among other tables, there are separate tables for agreement and inflection information, complementation patterns, spelling variants, and abbreviations and acronyms and their fully expanded forms.

Obtaining the SPECIALIST Lexicon:

The SPECIALIST lexicon is available as an open source resource as part of the SPECIALIST NLP tools. Distribution is subject to the terms and conditions specified in the distribution package.

Down load: **http://SPECIALIST.nlm.nih.gov/SpecialistLexicon.html**

Terms and Conditions: **http://SPECIALIST.nlm.nih.gov/TermsAndConditions**