Multi-lingual Features of the Unified Medical Language System

Olivier Bodenreider

Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA
obodenreider@mail.nih.gov

The Unified Medical Language System (UMLS) is a terminology integration system developed and maintained by the U.S. National Library of Medicine (NLM). Over the past 20 years, the UMLS Metathesaurus has been extended to encompass 168 source vocabularies. While English is the dominant language (116 source vocabularies), 52 vocabularies in the Metathesaurus are in languages other than English (6 in Dutch, German and Spanish ; 5 in French ; 4 in Italian and Portuguese ; 2 in Czech, Finnish, Hungarian, Japanese, Korean, Norwegian and Swedish ; and 1 in Basque, Croatian, Danish, Hebrew, Latvian, Polish and Russian).

NLM does not translate any of these vocabularies, but rather integrates translations of English vocabularies when they are available. Examples include the Medical Subject Headings (MeSH) present in the Metathesaurus in 16 languages ; the International Classification of Primary Care (ICPC) in 14 ; the Medical Dictionary for Regulatory Activities Terminology (MedDRA) in 10 ; and the WHO Adverse Drug Reaction Terminology in 5. Of note, SNOMED CT and the Physicians' Current Procedural Terminology (CPT) are both present in English and Spanish, because these two languages are the main languages used in the U.S.

Synonymous terms from source vocabularies are clustered into a UMLS concept regardless of language. This feature of the Metathesaurus makes it easy for users to identify non-English synonyms for a given concept and, conversely, to identify a concept from a term in a given language. Ignoring obsolete concepts, the Metathesaurus contains 2.8M concepts with English terms. With terms in 12% of the concepts, Spanish is the second most represented language in the Metathesaurus (mainly due to the presence of the Spanish translation of SNOMED CT). Other languages are represented in at most 3% of the concepts. Separate word indexes are provided for each language. The Metathesaurus has adopted the Unicode character set over 10 years ago and uses UTF-8 for character encoding.

UMLS users have taken advantage of the parallel corpus provided by the integration of vocabularies in multiple languages. Cross-lingual information retrieval systems such as BabelMeSH and MorphoSaurus leverage the UMLS Metathesaurus, which has also been used to support the translation of vocabularies such as SNOMED CT and the Foundational Model of Anatomy, and for acquiring parallel lexicons.

The UMLS is available at https://uts.nlm.nih.gov/. Using the UMLS Metathesaurus requires a free license, due to intellectual property restrictions with some of the source vocabularies.