UK SNOMED CT Lexicon

Published October 2018

Purpose of this Document

The purpose of this document is to provide background information for using the UK SNOMED CT Lexicon

The final UK SNOMED CT Edition release in Release Format 1 (RF1) was April 2018. As the Lexicon is not a supported product it is not considered for conversion in to Release Format 2 (RF2) and is not provided as part of the RF2 release. The April 2018 lexicon data is available within this document, (see Appendix 3) and there are no further plans to update this document after the October 2018 release.

For more information please contact the Information Standards Service Desk

Contents

1. About the UK SNOMED CT lexicon

As defined in the Oxford English Dictionary, a lexicon is “the vocabulary of a person, language, or branch of knowledge”. The lexicon is a word list with comprehensive coverage in both general English and clinical terms. It is actively updated and reviewed with new additions from the development of SNOMED CT® terminology. This is maintained and distributed by NHS Digital as the UK National Release Centre (UK NRC) for SNOMED CT.

1.1 Scope and usage

The lexicon plays an important role in the quality assurance of clinical terminology content. As a repository of correctly spelt words, the scope of the lexicon is those words which are acceptable in the UK edition of SNOMED CT. It is highly recommended that users understand the limitations and benefits of the lexicon to ensure appropriate implementation. The criteria for inclusion or exclusion of content in the lexicon are explained in the following section.

2. Development of the lexicon

A lexicon is a useful tool for assisting the manual efforts of the terminology authoring team in ensuring all content of the UK edition has been correctly spelt. Most commercially available dictionaries or lexicons do not include sufficient medical terms, whereas medical dictionaries lack coverage of general English. Furthermore, proprietary products limit the freedom of content control by users.

Originally a lexicon was developed for SNOMED CT based on the lexicon used for assurance of Read codes. The limitations of this approach were identified and it was decided to use an open source international development called the SPECIALIST lexicon as the baseline for UK specific lexicon for SNOMED CT.

2.1 Specialist lexicon as the baseline

The SPECIALIST lexicon is used as the baseline for the lexicon because it is a large syntactic lexicon of biomedical and general English. The sources for medical terms in SPECIALIST are:

The NLM Test Collection of MEDLINE abstracts

The UMLS Metathesaurus

Dorland’s Illustrated Medical Dictionary

Webster’s Medical Desk Dictionary

The sources for general English vocabularies in SPECIALIST are:

The American Dictionary Word Frequency Book

Longman’s Dictionary of Contemporary English

The Oxford Advanced Learner’s Dictionary

SPECIALIST is designed to provide the lexical information needed for the SPECIALIST Natural Language Processing (NLP) system (see appendix 1). It includes both US and UK English alternative spelling variations. The structure of the lexicon includes metadata, such as: category, spelling variant, indication of acronym, abbreviation or truncation etc. The content of the lexicon includes single words and word phrases or multiple-word terms, as well as hyphenated words.

SPECIALIST lexicon has been developed by the Lexical Systems Group of the Lister Hill National Centre for Biomedical Communications and released annually as one of the UMLS Knowledge Sources since 1994. It is available as an open source resource subject to terms and conditions (Appendix 1). Further detail on SPECIALIST can be found via the following link: http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicon.html

2.2 Inclusion and exclusion criteria for the lexicon

The comprehensive coverage of the open source SPECIALIST lexicon is an ideal baseline for the UK SNOMED CT lexicon, however it requires a degree of localisation for NHS purposes. The content of SPECIALIST has been modified accordingly. Inclusion and exclusion criteria have been applied for all new additions for the lexicon though they can be extended or adapted to meet the needs of new patterns.

2.2.1 Criteria for inclusion

The lexicon is currently used purely for spell checking in UK SNOMED CT terminology development as one component of routine quality assurance processes. The lexicon allows the inclusion of most correctly spelt words in modern English. Each criterion for inclusion is based on terminology editorial principles. The lexicon does not contain any linguistic information about a word itself such as derivation or pronunciation. These criteria can be expressed as:

- English words (UK spelling only)
- Clinical terms from other languages (mainly but not exclusively Latin and Greek)
- Proper names (e.g. Eponyms, place, religion, language, country, city)
- Both common and formal taxonomic names of organism (e.g. Animals and plants), chemicals, non-proprietary generic drug names
- Different forms of word (e.g. Plural, noun, adjective, tenses of verbs, pronoun)

The UK spelling variation is required to support NHS applications. Other variations such as US spellings are excluded from the UK lexicon for the national release. For example, oesophagus and haemodialysis are kept while esophagus and hemodialysis are excluded.

Clinical terms originating from Latin and Greek are commonly used today and form an important part of a clinical domain lexicon.

Inevitably, a very small number of words from other languages have been included in the UK lexicon from proper names, such as, the word ’van’ in the name ’van der Mee’.

Words containing non-English letters are not included. For example, protégé is not included because it contains accented letters. However, the accepted correctly spelt de-accented variation is included.

2.2.2 Criteria for exclusion

The following are not included in the UK SNOMED CT lexicon:

- Acronyms or abbreviations
- Proprietary or brand names (includes proprietary drug names)
- Phrases or multi-word terms (e.g. “myocardial infarction”)
- Hyphenated words
- Truncations
- Words with apostrophes
- Single letters

Acronyms and abbreviations are excluded to minimise ’false negative’ tests, e.g. A wrongly spelt word not being reported when it is coincidentally the same as an acronym or abbreviation.

2.3 Construction of the lexicon

The SPECIALIST lexicon released in 2009 was imported into the UK NRC terminology server without metadata and structures and the following modification steps taken:

1. Take the file LRWD.txt, which is the full single word index, and extract the first field, which contains the words themselves.
2. Remove all duplicates from this word list.
3. Remove any word not consisting entirely of alphabetic characters.
4. Take the file LRABR.txt, which is the abbreviations file, and extract field 2, the list of abbreviations.
5. Convert the abbreviations to lower case, and subtract them from the word list derived in step 3 (note that this removes a few genuine words, as some of the abbreviations are lexically valid in lower case).
6. Remove US spelling variations.
7. The remaining word list is the baseline for the UK lexicon.

Phrases and multi-word terms are not included in the lexicon but are broken down to single word entries, for example:

Multi-word term ’intermittent mandatory ventilation’ in SPECIALIST
Imported as three separated words: ’intermittent’, ’mandatory’, ’ventilation’

The Dutch name ’van der Mee’ in SPECIALIST
Imported as ’van’, ’der’, ’mee’ in the UK lexicon

Word truncations are not included in the UK lexicon, for example:

’A. Tuberosum’ is a truncation of ’Allium tuberosum’ in SPECIALIST
Imported as two separate words: ’allium’, ’tuberosum’ in the UK lexicon

Any word with apostrophe or hyphenated words are not included, for example:

’Morton's disease’ in SPECIALIST
Imported as ’morton’, ’disease’ in the UK lexicon

’Mosetig-Moorhof bone wax’ in SPECIALIST
Imported as ’mosetig’, ’moorhof’, ’bone’, ’wax’ in the UK lexicon

3. Maintenance and release of the lexicon

October 2011 was the first release of the lexicon. It was shipped with the main RF1 SNOMED CT Release under the folder:

SnomedCT_RF1Release_GB1000000_YYYYMMDD\Resources\UKLexicon

As previously described the final UK SNOMED CT Edition release in Release Format 1 (RF1) was April 2018. As the Lexicon is not a supported product it is not considered for conversion in to Release Format 2 (RF2) and is not provided as part of the RF2 release. The April 2018 lexicon data is available within this document, (see Appendix 3) and there are no further plans to update this document after the October 2018 release.

The lexicon was provided as a ‘technical preview’ in a tab delimited text file and was used primarily for spell checking in word repositories used in healthcare (see Appendix 2 for definition).

Each RF1 terminology release included an updated version of the lexicon, the final version is available in appendix 3.

Appendix 1

Terms and Conditions for Use of the SPECIALIST NLP Tools

1. Introduction

The following Terms and Conditions apply for use of the SPECIALIST NLP Tools. Using the SPECIALIST NLP Tools indicates your acceptance of the following Terms and Conditions. These Terms and Conditions apply to all SPECIALIST NLP Tools, independent of format and method of acquisition.

2. The SPECIALIST NLP Tools

The Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, has developed the SPECIALIST NLP Tools to investigate the contributions that natural language processing techniques can make to the task of mediating between the language of users and the language of online biomedical information resources. The SPECIALIST NLP Tools facilitate natural language processing by helping application developers with lexical variation and text analysis tasks in the biomedical domain.

3. Availability

The SPECIALIST NLP Tools are available to all requesters, both within and outside the United States, at no charge.

4. Use of the SPECIALIST NLP Tools

A. Redistributions of the SPECIALIST NLP Tools in source or binary form must include this list of conditions in the documentation and/or other materials provided with the distribution.

B. In any publication or distribution of all or any portion of the SPECIALIST NLP Tools (1) you must attribute the source of the tools as the SPECIALIST NLP Tools with the release number and date; (2) you must state any modifications made to the SPECIALIST NLP Tools along with a complete description of the modifications, which may be in the form of patch files.

C. You shall not assert any proprietary rights to any portion of the SPECIALIST NLP Tools, nor represent the SPECIALIST NLP Tools or any part thereof to anyone as other than a United States Government product.

D. The name of the U.S. Department of Health and Human Services, National Institutes of Health, National Library of Medicine, Lister Hill National Center for Biomedical Communications may not be used to endorse or promote products derived from the SPECIALIST NLP Tools without specific prior written permission.

E. Neither the United States Government, U.S. Department of Health and Human Services, National Institutes of Health, National Library of Medicine, Lister Hill National Center for Biomedical Communications, nor any of its agencies, contractors, subcontractors or employees of the United States Government make any warranties, expressed or implied, with respect to the SPECIALIST NLP Tools, and, furthermore, assume no liability for any party's use, or the results of such use, of any part of these tools.

These terms and conditions are in effect as long as the user retains any part of the SPECIALIST NLP Tools.

Appendix 2

Interpretation of materials distributed as ‘technology preview’:

1. The release format specification of the product is public but not fixed.

2. The method of product content preparation may be public (but is not required to be) but is not fixed.

3. Quality or safety assurance of the product may be ill-defined and/or incomplete.

4. So far as is possible within internal resourcing constraints, UK NRC undertakes to support the product and maintain it in synchronisation with the primary release data as required.

5. Trial implementation is encouraged to evaluate utility and safety, and to identify possible design improvements.

6. Live deployment is at the users own risk, where such risk is permitted by other governance processes.

7. Implementations may have to change if and when the product design changes in the light of experience.

8. In the event that the product is found to be only minimally used, not useful, or unsafe, then the UK NRC is under no obligation to continue its support or maintenance.

Appendix 3

Lexicon from the final UK SNOMED CT RF1 release in April 2018

lexicon_20180401.txt

Information Standards Service Desk
Tel: 0300 30 34 777
E-mail:
information.standards@nhs.net
Internet:
https://digital.nhs.uk/snomed-ct
Terminology and Classifications Delivery Service
NHS Digital
1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE