# Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping

**Chris J. Lu[a, b], Destinee Tormey[a, b], Lynn McCreedy[a, b] and Allen C. Browne[a]**

[a.] *National Library of Medicine, Bethesda, MD, USA*
[b.] *Medical Science & Computing, Inc., Rockville, MD, USA*

**Abstract**

*Concept mapping is important in natural language processing (NLP) for bioinformatics. The UMLS Metathesaurus provides a rich synonym thesaurus and is a popular resource for concept mapping. Query expansion using synonyms for subterm substitutions is an effective technique to increase recall for UMLS concept mapping. Synonyms used to substitute subterms are called element synonyms. The completeness and quality of both element synonyms and the UMLS synonym thesaurus is the key to success in such applications. The Lexical Systems Group (LSG) has developed a new system for element synonym acquisition based on new enhanced requirements and design for better performance. The results show: 1) A growth of 36.71 times of synonyms in the Lexicon (lexSynonym) in 2017 release; 2) improvements of concept mapping on recall and F1 with similar precision using the lexSynonym.2017 as element synonyms due to the broader coverage and better quality.*

***Keywords:***

Synonyms, Natural Language Processing, Query Expansion, The UMLS SPECIALIST Lexicon, UMLS Conept Mapping

## Introduction

Subterm substitution is a popular technique in query expansion. It is used to increase recall when no direct UMLS concept mapping is found through normalization. For example, no concept is found by direct mapping through normalization if the source vocabulary is "*nasal deformity*". By substituting the subterm "*nasal*" for its synonym, "*nose*", the UMLS concept [C0240547, Nose Deformity] is found, where "C0240547" is the concept unique identifier (CUI) and "*Nose Deformity*" is the preferred term in the UMLS. In this example, "*nasal*" and "*nose*", which are used for substitution, are called element synonyms, while "*nasal deformity*" and "*nose deformity*" are the input term and expanded term, as shown in Figure 1. The normalized form of expanded terms is then used for concept mapping from UMLS synonyms.

Element synonyms are semantically equivalent terms (e.g. "*nasal*" and "*nose*" in the example above) used to identify subterms in the source vocabulary for substitution in UMLS concept mapping. This method increases recall by finding concepts for terms whose concept cannot found by normalization or not even in the UMLS. For example, if "*elderly*" and "*geriatric*" are element synonyms, "*elderly patients*", a term in the corpus (PubMed) but not in the UMLS, is mapped to the UMLS concept [C0199167, geriatric patients] by substituting "*elderly*" with its synonym, "*geriatric*". The performance of this method relies on the quality and completeness of the element synonyms for a given UMLS thesaurus. The broader the coverage of the element synonyms, the higher the recall. Commutativity and transitivity are two needed properties for quality element synonyms to preserve precision.

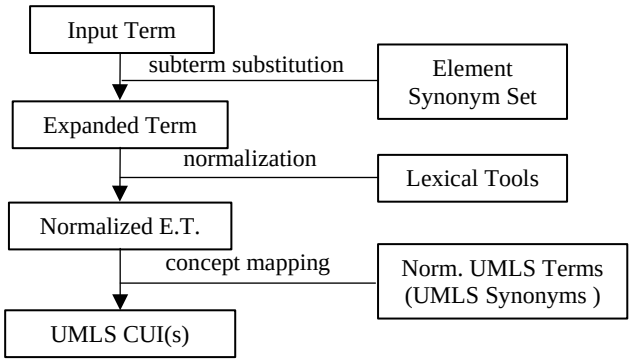


*Figure 1 – Element Synonyms and Subterm Substitution in UMLS Concept Mapping*

In this paper, we present a systematic approach to acquire a set of high quality element synonyms from the SPECIALIST Lexicon and UMLS Metathesaurus. The results show an improvement on recall and F1 with similar precision using this new acquired element synonym set for concept mapping.

## Background

The 2016AA UMLS Metathesaurus of the National Library of Medicine (NLM), containing more than 3.25 million concepts and nearly 13 million unique concept names from over 190 source vocabularies, is one of the richest thesauri in the biomedical domain. UMLS concept mapping is used for managing knowledge in NLP applications including information retrieval (IR), document retrieval (DR), text classification, data mining, and decision support systems. Normalization is used as the initial step for UMLS concept mapping. All UMLS terms are processed through the Norm program in the Lexical Tools to normalize lexical variants, syntactic representation, and character encoding between ASCII and Unicode [1-2]. For example, "*Behcet diseases*", "*Behcet's disease, nos*", and "*disease, Behçet*" are UMLS synonyms because they represent the same concept. They have the same normalized term "*behcet disease*". All UMLS terms are normalized and stored in the UMLS (MRXNS_ENG.RRF) with their associated concept(s). Terms having the same normalized form from input vocabulary (even if they are not in the UMLS Metathesaurus) can be mapped to UMLS concepts. For example, "*disease, Behcet*", which is not a UMLS term, like other terms above, is mapped to [C0004943, Behcet Syndrome] through this normalization process.

Subterm substitution is used to find concepts for terms whose concept cannot be found through normalization. To increase

recall, strategies may use lexical or semantic information, or a combination of both. First, subterms can be substituted by lexically related variants, such as derivations. Derivations allow users to find closely related terms that may differ by part of speech (POS) for better recall [3-4]. For example, no CUI is found by direct mapping through normalization if the source vocabulary is "*perforated ear drum*". By substituting the subterm "*perforated*" for its derivational variant, "*perforation*," the UMLS concept is found [C0206504, Tympanic Membrane Perforation]. Second, subterm substitution by semantically equivalent terms (synonyms) improves recall [5-6]. Synonyms used for subterm substitution are terms that have the same meaning (concept) and are called element synonyms (e.g. "*nasal*" and "*nose*" in the example above). In practice, synonyms of synonyms are retrieved recursively (recursive synonyms) in such applications to increase recall. Third, subterms can be substituted by a combination of both lexical variants and synonyms [7-9]. These applications usually pre-generate all expanded terms and use them in a pool for concept mapping. The broader the coverage of the expanded terms, the better the recall for such approaches. Several works used this strategy to find terms that the UMLS missed to improve recall [10-11]. This method of subterm substitutions generates many mapped concepts, including irrelevant concepts, and results in higher recall and lower precision. Ranking and filters, such as keyword match, frequency (TF-IDF), semantic types, concept distance and the longest lead-terms or end-terms, are used to improve the precision [12-13]. Other research has focused on different query expansion strategies by using UMLS Tools [14-15], MeSH [16-17] or their application systems [18-19] for effective UMLS concept mapping and information retrieval. Some research has explored the role of semantic similarity and semantic relatedness to similar and related terms having different UMLS concepts [20-21]. Prior to our work, there has been very limited effort devoted to acquiring element synonyms. Synonyms in the UMLS and Lexicon are two of most commonly used sources for element synonyms. However, several issues are found as described below.

UMLS synonyms with some restrictions, such as source vocabulary (MeSH), term length, and size of grams (usually unigram), were used as element synonyms for UMLS concept mapping in previous research [7-11, 14-16]. Three issues have been found in such approaches. First, UMLS synonyms are over-generated for element synonyms. For example, "*allergy drug*" and "*allergy medicine*" are UMLS synonyms, [C0013182, Drug Allergy] and considered as expanded terms. The concepts of these expanded terms can be found if their subterm, "*drug*" and "*medicine*", are in an element synonym set. Slow runtime performance and computer resources are other concerns in practice when using the expanded terms of UMLS synonyms as element synonyms in subterm substitution, due to the large-scale size. Second, element synonyms must have properties of commutativity and transitivity for effective concept mapping. For example, "*ago*" is the abbreviation (ISO country code) for the country "*Angola*" and thus they are UMLS synonyms (with the same CUI, C0003023). However, "*ago*" is more often associated with another meaning, 'earlier,' and is not a synonym for "*Angola*" (lack of commutativity). In short, UMLS synonyms that represent broader or narrower concepts (such as "*adnexa*" and "*uterine adnexa*"), acronyms, abbreviations, POS ambiguity (e.g. "*mushroom*" is a synonym of "*Agaricales*" when its POS is a noun, but the meaning shifts when its POS changes to a verb), terms with multiple CUIs, or the combination of above, should be excluded from element synonyms. Acronyms with multiple CUIs cause steep precision drop by a huge amount of irrelevant mapped concepts in recursive subterm substitutions. For example, the acronym of "*ER*" has more than 27 different expansions (concepts), such as "*emergency room*", "*efficacy ratio*", "*eye research*", etc. Third, element synonyms may be single words (unigrams) or multiwords (words with spaces). Terms (multiwords) are used in more sophisticated systems as element synonyms to gain better recall [5-11]. For example, no concept is found for "*zona vaccine*". By substituting "*zona*" for its multiword synonym, "*herpes zoster*", the UMLS concept (C1720918) is found. On the other hand, longer terms introduce more noise rather than improving the performance. For example, "*herpes zoster infection*" a UMLS synonym of "*zona*" should not be used as an element synonym. To our best knowledge, there is no study on how many grams should be used for element synonyms.

In addition to UMLS synonyms, LexSynonyms are also commonly used as element synonyms for UMLS concept mapping in NLP. They are recorded in the format of synonym pairs (sPairs) with POS information and distributed with the SPECIALIST Lexicon. Two synonym records (sRecords) are generated by an sPair because sPairs are bi-directional. They are in the format of [synonym-1|POS-1|synonym-2|POS-2]. In most applications, they are integrated with lexical variants, to generate expanded equivalent terms (fruitful variants) for concept mapping in MetaMap [7], MMTx [8], and Sophia [9]. LexSynonyms were originally collected as a set in the early 1990s and maintained manually by LSG linguists based on users' requests. A rather static size of this synonym set is observed: only 142 sRecords were added between 2004 (5,056) and 2016 (5,198). Thus, we developed a systematic approach to acquire lexSynonyms as a standalone set of element synonyms with greater coverage and better quality for more effective UMLS concept mapping and NLP applications that use synonym retrieval.

## Approaches

Synonyms can be categorized into two types: cognitive synonyms and near-synonyms. Cognitive synonyms have fewer meaning differences with greater interchangeability, while near-synonyms lack these. Cognitive synonyms match the characteristics of element synonyms well for effective performance (recall and precision) because they have two properties, commutativity and transitivity. Commutativity, (x = y) -> (y = x), preserves the naturalness of bi-direction of sPairs. For example, if "*joy*" is a cognitive synonym of "*happy*", then "*happy*" is a cognitive synonym of "*joy*". Transitivity, ((x = y) and (y = z)) -> (x = z), preserves the precision in recursive synonym applications. For example, if "*happy*" is a synonym of "*joy*", and "*joy*" is a synonym of "*enjoy*", then "*happy*" is a synonym of "*enjoy*". These two properties are necessary conditions of quality element synonyms for subterm substitutions in concept mapping. However, they are missing in most synonym sets used in NLP. They are required for lexSynonym acquisition in our new system to ensure the effective UMLS concept mapping: lexSynonyms must be cognitive synonyms.

To acquire a thorough synonym set, UMLS synonyms are chosen as source candidates in this project. UMLS synonyms are UMLS strings (element terms and expanded terms) with the same concept (CUI). They are grouped and represented as a key-value collection in a synonym class (sClass). Namely, the key is the CUI while the value is the list of all terms with the same CUI in the UMLS Metathesaurus. This is the common way of retrieving UMLS synonyms. The derived UMLS sClass is further enhanced through the integration of the Lexicon. The Lexicon includes additional information needed for resolving the NLP issues mentioned above, such as POS, inflections, acronyms, abbreviations, etc. First, a lexical entry must be a word (single word or multiword) with a special unit of meaning

in itself [22-23]. The Lexicon is used as the source vocabulary to filter element synonyms: terms in the sClass that are not in the Lexicon, such as non-word phrases, are removed. For example, expanded terms of UMLS synonyms "*allergy drug*" and "*allergy medicine*" are removed to resolve the issue of over-generation, while "*herpes zoster infection*" is removed to resolve the issue of n-grams because none of them are in the Lexicon (do not meet the requirements of LexMultiwords) [23]. As discussed before, recall of concept mapping will not decrease because, "*drug*" and "*medicine*"; "*zona*" and "*herpes zoster*", are terms in the Lexicon and used as element synonyms. Second, the POS information from the Lexicon is added to the sClass to resolve the POS ambiguity issues. Third, terms that are acronyms or abbreviations in the Lexicon are removed to preserve precision. Fourth, synonyms in the sClass need to be verified by experts to ensure they meet the requirements of commutativity and transitivity. Finally, the verified sClass is further processed into sPairs and sRecords to compose the element synonym set. All synonymous terms from the Lexicon (lexSynonyms) are acquired using this approach.

## Implementation

A standalone lexSynonym set is established by collecting all synonymous terms in the Lexicon based on the above requirements and approaches. LexSynonyms are acquired from three types of sources: the Lexicon, the UMLS, and NLP projects. They are described as follows.

***Lexicon-Sourced Synonyms – Nominalizations with EUI***

Nominalizations are cognitive synonyms with the adjectives and/or verbs from which they are derived. They are recorded in the Lexicon and can be retrieved automatically to generate lexSynonyms. Additional information, the entry unique identifier (EUI) of the lexical record, is added to the associated sPair for downstream NLP processing. For example, the sPair of [ability|noun|able|adj|E0006490] is generated from the lexical record (E0006490). As shown in Figure 2, the noun of "*ability*" is the nominalization of the adjective, "*able*".

```
{base=ability
entry=E0006490
        cat=noun
        variants=reg
        variants=uncount
        compl=pphr(of,np)
        compl=infcomp:arbc
        nominalization_of=able|adj|E0006510
}
```

*Figure 2 – Lexical Record of C0011065, ability*

***UMLS-Sourced Cognitive Synonyms with CUI***

The Lexicon and UMLS Metathesaurus are used to retrieve more synonymous lexicon terms as follows. First, all English terms from the UMLS (MRCONSO.RRF) with the same EUI are retrieved. Second, concepts of chemicals and drugs are removed due to limited resources and application domains. The semantic type indexes (STIs) of chemicals and drugs are used as filters through the mapping from CUI to STI (MRSTY.RRF). Third, terms having the POS of noun, verb and adjective with inflections of base in the Lexicon are retrieved. This step eliminates inflectional variants, illegal POSs, and non-word phrases from the UMLS synonyms. Fourth, terms that are acronyms or abbreviations in the Lexicon are removed. Fifth, terms with the same CUI are stored in an sClass with the CUI as the key and a list of terms as the value. The associated EUI is added to each term in the list for the computer to reference lexical records for needed information. Sixth, terms that are spelling variants (spVar) or nominalizations of other terms in the same sClass are removed to save manual tagging time because they can be generated automatically later (in step nine). Seventh, sClasses with only one term are removed because they do not have synonyms. Eighth, UMLS preferred terms are added to sClasses for concept identification by LSG linguists when validating if terms (synonym candidates) are cognitive synonyms of the sClass. Ninth, spVars and nominalizations of validated synonym candidates are added back to the sClass. Tenth, tagged sClasses are used to generate the sPairs and sRecords with POS and source information (CUI, EUI and NLP). For example, "*death*", "*dead*", "*deceased*" and "*die*" are base forms with qualified POSs in the Lexicon, have the same CUI (C0011065), and are not chemicals, drugs, acronyms, or abbreviations. They are thus synonym candidates and are gathered in a candidate sClass as shown in Figure 3. Among the synonyms, "*die*" is related by nominalization to "*death*" (E0020918), and is thus removed. This is the candidate sClass sent to LSG linguists for validation. Cognitive synonyms are tagged as "Y" while near-synonyms are tagged as "N". The nominalizations, "*deadness*" (E0020885) from "*dead*" (E0020877) and "*die*" from "*death*", are added back into the sClass automatically. The final sClass is composed of 5 synonyms, generating 10 (bidirectional) sPairs, and results in 20 synonym records (sRecords) in the lexSynonym set, as shown in Figures 4 and 5 respectively.

```
#SYNONYM_CLASS|C0011065|Cessation of life
noun|E0020918|death|
adj|E0020877|dead|
adj|E0020990|deceased|
verb|E0022536|die|
```

*Figure 3 – Example of Candidate sClass: C0011065*

```
#SYNONYM_CLASS|C0011065|Cessation of life
noun|E0020918|death|Y
adj|E0020877|dead|Y
adj|E0020990|deceased|Y
verb|E0022536|die|nom
noun|E0020885|deadness|nom
```

*Figure 4 – Example of Final sClass: C0011065*

```
deadness|noun|dead|adj|C0011065
deadness|noun|death|noun|C0011065
deadness|noun|deceased|adj|C0011065
deadness|noun|die|verb|C0011065
dead|adj|deadness|noun|C0011065
dead|adj|death|noun|C0011065
dead|adj|deceased|adj|C0011065
dead|adj|die|verb|C0011065
death|noun|deadness|noun|C0011065
death|noun|dead|adj|C0011065
death|noun|deceased|adj|C0011065
death|noun|die|verb|C0011065
deceased|adj|deadness|noun|C0011065
deceased|adj|dead|adj|C0011065
deceased|adj|death|noun|C0011065
deceased|adj|die|verb|C0011065
die|verb|deadness|noun|C0011065
die|verb|dead|adj|C0011065
die|verb|death|noun|C0011065
die|verb|deceased|adj|C0011065
```

*Figure 5 – Example of sRecords: C0011065*

***NLP Project-Sourced Cognitive Synonyms***

Synonyms from NLP projects can be processed by similar steps to those described above, then added into lexSynonyms. For the 2017 release, we processed synonyms from Lexical Variants Generation (LVG). Duplicated synonyms of the previous two sources are removed from the candidate list without further process. Others are converted to sPair candidates computationally, reviewed by LSG linguists, and added to the lexSynonym set with POS if they are cognitive sPairs and in the Lexicon. "NLP_XXX" is used as the source information for the NLP project "XXX". For example, "NLP_LVG" is marked as the source for synonyms from the LVG. The NLP project-

sourced synonyms provide two important features of extendibility and compatibility. First, users are able to extend the synonym set by adding domain/project specific synonyms. Second, it preserves the same result for the specific NLP project (LVG) users when forward compatibility is required.

## Results, Tests, Discussions and Applications

As a result, 22,779 sClasses and 58,134 synonym candidates are retrieved from the UMLS source type (2016 AA UMLS Metathesaurus and 2016 Lexicon). Cognitive synonyms from this candidate list are used to generate 118,468 sRecords. In addition, 67,584 sRecords from Lexicon nominalizations and 4,792 sRecords from NLP_LVG are generated, respectively. All sRecords from these resources are combined into the lexSynonym set and distributed in the 2017 release of the Lexicon. The results show a growth of 36.71 times from 2016 to 2017 release through this new approach (Table 1).

*Table 1 – Growth for LexSynonyms 2016 to 2017*

| Year | CUI | EUI | NLP | Total |
|---|---|---|---|---|
| 2016 | 0 | 0 | 5,198 | 5,198 |
| 2017 | 118,468 | 67,584 | 4,792 | 190,844 |

A model is established to measure the performance of using the lexSynonym.2017 for UMLS concept mapping through the Sub-Term Mapping Tools (STMT). STMT applies a real-time subterm substitution algorithm for UMLS concept mapping with configurable options of choosing element synonyms and UMLS release. The UMLS-CORE project assigned CUI(s) to terms (13,076) that are within the top 95% usage and mappable to SNOMED CT [5]. 2,755 of these terms (with 2,756 CUIs) without mapped concepts in UMLS.2016AB through normalization are used as the gold standard for this test. Five normalized element synonym sets are configured in STMT for comparison. The default STMT element synonym set is comprised of high quality synonyms for subterm substitution to improve recall (25%). They are validated cognitive synonyms from sources of British English, Greco-Latin, acronyms, abbreviations, Emergency Care Research Institute (ECRI), etc. [6]. Results are shown in Table 2: 1) recall is increased over 10% from lexSynonym 2016 to 2017 due to broader coverage (from 5K to 150K). Also, the precision is increased due to better quality. 2) recall and F1 are further improved about 5% and 0.05 while precision is about the same (-0.03%) by adding 2017 lexSynonyms to the STMT synonym set. The set of lexSynonym.2017 contains 5,872 (~75%) normalized synonyms in the STMT synonym set. Adding the previous lexSynonyms (2016) to STMT offers no improvement.

*Table 2 – Test Result for Terms without Mapped Concepts*

| Synonym Set | N. Size** | Prec. | Recall | F1 |
|---|---|---|---|---|
| STMT | 7,873 | 66.16% | 25.04% | 0.3633 |
| LS.2016* | 5,070 | 42.86% | 0.33% | 0.0065 |
| LS.2017 | 149,912 | 71.04% | 10.41% | 0.1816 |
| STMT+LS.2016 | 12,681 | 65.87% | 25.07% | 0.3632 |
| STMT+LS.2017 | 151,913 | 66.13% | 30.04% | 0.4132 |

* LS: LexSynonym Set, **N.: Size of Normalized Synonym Set

Due to limited resources, about 1/3 of synonym candidates (20,566 out of 58,134) have so far been tagged. The properties of commutativity and transitivity of lexSynonyms are ensured by nominalization (Lexicon-sourced) or by linguists' tags. 92.20% of synonym candidates are tagged as "Y". The size of the UMLS-sourced lexSynonym is about 0.64% of the size of the UMLS synonyms in English. Accordingly, the size of lexSynonyms will be about 2% of the UMLS synonyms when the tagging process is completed. LexSynonyms thus yield a much smaller, more manageable set to be used as element synonyms. In addition, synonyms from other NLP projects, such as UMLS-CORE and STMT, can be further processed and added to the lexSynonyms. Recall is expected to be further improved as the size of element synonyms grows while the precision is preserved by the properties of cognitive synonyms.

We utilized lexSynonyms as element synonyms in NLP applications (Lexical Tools) to retrieve synonyms. Synonyms, POS, and source information are provided in the outputs of synonym features of Lexical Tools. A sophisticated algorithm is implemented as follows in the recursive synonym flow component to preserve precision. First, only synonyms with the same CUI are retrieved recursively if the source type is CUI. Second, all synonyms are retrieved recursively if the source type is EUI. Third, synonyms from the same NLP projects are retrieved recursively if the source type is NLP. In addition, the synonym source option (-ks) is implemented to allow users to restrict the results by source type (CUI, EUI, NLP), or any combination of the above. These new features provide needed information to preserve precision for downstream NLP processing. For example, the five synonyms of "*die*" are retrieved from the synonym feature (-f:y) in Lexical Tools. The source information is also included. As shown in Figure 6, "*dead*", "*deadness*", "*death*" and "*deceased*" are from the source of UMLS with CUI of [C0011065], while "*expire*" is from source of NLP (project LVG). The POS information is included in the outputs of the Lexical Tools. "*Terminate*", a synonym of "*expire*" from the resource of NLP_LVG, is retrieved when the recursive synonym feature (-f:r) is used in the Lexical Tools, as shown in Figure 7. The last two fields of the last line in Figure 7 show the source type (NLP_LVG) and the recursive history (yy, means synonym of synonym). Thus, project specific non-cognitive sPairs, "*dead*" and "*terminate*", can be distinguished by the different types of sources (CUI vs NLP) to preserve the precision in recursive synonyms.

```
die|verb|y|dead|adj|C0011065
die|verb|y|deadness|noun|C0011065
die|verb|y|death|noun|C0011065
die|verb|y|deceased|adj|C0011065
die|verb|y|expire|verb|NLP_LVG
```

*Figure 6 –Synonyms of "die" from Lexical Tools*

```
die|verb|r|dead|adj|C0011065|y
die|verb|r|deadness|noun|C0011065|y
die|verb|r|death|noun|C0011065|y
die|verb|r|deceased|adj|C0011065|y
die|verb|r|expire|verb|NLP_LVG|y
die|verb|r|terminate|verb|NLP_LVG|yy
```

*Figure 7 – Recursive Synonyms of "die" from Lexical Tools*

## Conclusion

We have demonstrated the usefulness of the general concept of element synonyms as well as the Lexicon-specific type of element synonyms, lexSynonyms, in concept mapping. A systematic and maintainable approach is used to acquire higher quality lexSynonyms through the use of the Lexicon. Issues of over-generation and n-grams are resolved by restricting UMLS synonyms that are base forms with noun, verb, and adjective POS in the Lexicon, and removing chemicals and drugs. Terms that are acronyms or abbreviations are removed to avoid a drop in precision. Synonym candidates in the sClass that do not match the properties of commutativity and transitivity are tagged by the linguists as invalid to resolve near-synonym issues. POS is added to sPairs automatically through a Lexical records mapping by using EUIs in the sClass during the generation process. The information of source with unique identifier (CUI, EUI, and NLP) is also included. This

information is vital for downstream NLP applications to preserve precision especially when recursive synonyms are used. As a result, a thorough set of element synonyms is generated. LexSynonyms are expected to grow with the Lexicon and UMLS Metathesaurus for better coverage through this system. This approach is generic for element synonym acquisition and can be applied to other corpora, vocabularies, or synonym thesauri. The generated lexSynonyms are used in the Lexical Tools with enhanced recursive algorithms to provide better usage of the synonym related features for NLP applications. We believe the impact of better quality and broader coverage for lexSynonym acquisition in the Lexicon for effective UMLS concept mapping will improve the precision, recall, and naturalness of NLP applications. The set of lexSynonyms is distributed in the 2017 release of SPECIALIST Lexicon with UMLS by NLM via an Open Source License agreement.

## Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would like to thank Dr. Kin Wah Fung, Dr. Marcelo Fiszman, Guy Divita, Willie Rogers, James Mork and Francois-Michel Lang for their valuable discussions and suggestions.

## References

[1] A.T. McCray, S. Srinivasan, A.C. Browne. Lexical Methods for Managing Variation in Biomedical Terminologies. In proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994, 235-239.

[2] C.J. Lu and A.C. Browne. Converting Unicode Lexicon and Lexical Tools for ASII NLP Applications. In proceedings of AMIA Annual Symposium, Oct. 22-26, 2011, 1870.

[3] C.J. Lu, D. Tormey, L. McCreedy, A.C. Browne. A Systematic Approach for Automatically Generating Derivational Variants in Lexical Tools Based on the SPECIALIST Lexicon. IEEE IT Professional Magazine, May/June, 2012, 36-42.

[4] C.J. Lu, D. Tormey, L. McCreedy, A.C. Browne. Generating SD-Rules in the SPECIALIST Lexical Tools - Optimization for Suffix Derivation Rule Set. HealthInf 2016, Feb. 21-23, 2016, Vol. (5), 353-358.

[5] K.W. Fung and J. Xu. An exploration of the properties of the CORE problem list subset and how it facilitates the implementation of SNOMED CT. JAMIA 2015, 22: 649-658.

[6] C.J. Lu, and A.C. Browne. Development of Sub-Term Mapping Tools (STMT). In Proceedings of AMIA Annual Symposium, Nov. 3-7, 2012, 1845.

[7] A.R. Aronson. The Effect of Texture Variation on Concept based Information Retrieval. In proceedings of AMIA Annual Symposium, 1996, 373-377.

[8] G. Divita, T. Tse, L. Roth. Failure Analysis of MetaMap Transfer (MMTx). In proceedings of MedInfo 2004, Sept. 7-11, 2004, 763-767.

[9] G. Divita, Q.T. Zeng, A.V. Gundlapalli, et al. Sophia: A Expedient UMLS Concept Extraction Annotator. In proceedings of AMIA Annual Symposium, Nov. 15-19, 2014, 467-476.

[10] W. Hole, S. Srinivasan. Discovering Missed Synonymy in a Large Concept-Oriented Metathesaurus. In proceedings of AMIA Annual Symposium, Nov. 4-8, 2000, 354-358.

[11] K.C. Huang, J. Geller, M. Halper, J.J. Cimino. Piecewise Synonyms for Enhanced UMLS Source Terminology Integration. In proceedings of AMIA Annual Symposium, Nov. 10-14, 2007, 339-343.

[12] T.C. Eskridge, A. Granados, A.J. Cañas. Ranking Concept Map Retrieval in the CmapTools Network. In proceedings of the 2nd International Conference on Concept Mapping, 2006. Vol. 1, 477-484.

[13] H.C. Wu and R.W.P. Luk. Interpreting TF-IDF Term Weights as Making Relevance Decisions. ACM Transactions on Information Systems, June, 2008, Vol. 26, No. 3, Article 13.

[14] N. Griffon, W. Chebil, L. Rollin, G. Kerdelhue, B. Thirion, J.F. Gehanno, and S. J. Darmoni. Performance evaluation of unified medical language system®'s synonyms expansion to query PubMed. BMC Medical Informatics and Decision Making 2012, 12(1): 12.

[15] K. Lu, X.M. Mu. Query Expansion Using UMLS Tools for Health Information Retrieval. J. of the American Society for Information Science and Technology, 2009, 46 (1), 1-16.

[16] M.C. Díaz-Galiano, M.T. Martín-Valdivia, L.A. Ureña-López. Query expansion with a medical ontology to improve a multimodal information. Comp. in Bio. and Med. 2009, 39(4): 396-403.

[17] M. Berardi, M. Lapi, P. Leo, and C. Loglisci. Mining Generalized Association Rules on Biomedical Literature. IEA/AIE, 2005, 500-509.

[18] P. Srinivasan. Query Expansion and MEDLINE. J. of Information Processing & Management. 1996, Vol. 32, No. 4, 431-443.

[19] Q.T. Zeng, D. Redd, T. Rindflesch, J. Nebeker. Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents. In proceedings of AMIA Annual Symposium, Nov. 3-7, 2012, 1050-1059.

[20] T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, C.G. Chute. Measures of Semantic Similarity and Relatedness in the Biomedical Domain. J. of Biomedical Informatics, 2006, 40(3), 288-299.

[21] S.V.S. Pakhomov, B. McInnes, T. Adam, Y Liu, T. Pedersen, G. Melton. Semantic Similarity and Relatedness between Clinical Terms: An Experimental study. In proceedings of AMIA Symposium, Nov. 13-17, 2010, 572-576.

[22] C.J. Lu, D. Tormey, L. McCreedy, A.C. Browne. Multiword Frequency Analysis Based on the MEDLINE N-Gram Set. In proceedings of AMIA Annual Symposium, USA, Nov. 12-16, 2016, 1488.

[23] C.J. Lu, D. Tormey, L. McCreedy, A.C. Browne. Generating A Distilled N-Gram Set: Effective Lexical Multiword Building in the SPECIALIST Lexicon. HealthInf 2017, Porto, Portugal, Feb. 21-23, 2017, Vol. 5, 77-87.

**Address for correspondence**

Dr. Chris J. Lu,

chlu@mail.nih.gov

NIH/NLM/LHC/CgSB/MSC

8600 Rockville Pike, Bldg. 38-A, B1N-28R,

Bethesda, MD 20894