Rules for Lexicon Terms

Lynn McCreedy
Destinee Tormey
Amanda Payne (2019-)

As a scientific field, biomedicine's concepts are ever-expanding, and with them come new terms. Our aim is to discern these terms and make records for them in a principled and rigorous way. Our principles derive from those set out at the Lexicon's inception by Allen Browne, Alexa McCray and others, as shown here in excerpts from the 1995 manual:

> **1.1 General Description:**
>
> **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.**
>
> **1.2 The Scope of the Lexicon**
>
> **Words are selected for lexical coding from a variety of sources. [W]ords from…MEDLINE abstracts together with words which appear both in the UMLS Metathesaurus and Dorland's …form the core of the words entered. In addition, an effort was made to include words from the general English vocabulary. [Goes on to list sources for general English, as well as verbs and adjectives used in biomedical English.]**
>
> **A variety of reference sources are used in coding lexical records. Coding is based on actual usage in the UMLS Test Collection, dictionaries of general English, primarily learner's dictionaries which record the kind of syntactic information needed for NLP, and medical dictionaries.**

So, the guidance we have taken from that, is that the Lexicon entries we make should:

- Be useful for NLP.
- Reflect biomedical vocabulary and usage (most general English terms are already in the Lexicon), as shown in published medical dictionaries and edited biomedical usage (MEDLINE, the Metathesaurus, Google Scholar, published works).

That is, we look for terms of art in biomedicine. The variants (spelling variants, grammatical variants) along with the complement information, nominalization, etc., which we include in each record are based on edited biomedical usage available to us. We have found Essie Hit Patterns to be key in this analysis, because this resource tells us frequency of occurrence in Medline, plus vital linguistic contextual information in patterns of preceding and following text.

Multiword expressions (LMWs, LexMultiWords) have been included in the Lexicon for more than 20 years. They have their roots in Lexbuild manual guidance, which in turn follows published medical dictionaries that include them, often as subterms within a more general term. An example from Dorland's: *law* has about 3 pages of subterms, beginning with *Allen's paradoxic law, all-or-none law, Angstrom's law, Arndt's law, Arndt-Schultz law, laws of articulation, Avogadro's law.* Dictionary subterms like these are examined as potential Lexbuild multiwords, but not all will qualify. Because "law(s) of articulation" is a noun with a postmodifying prepositional phrase, rather than being a single NP, it cannot be a Lexbuild base. Note that this restriction on complex NPs can be overridden if a term is a true compound, with its own meaning apart from the constituent NP + PP. For instance, "tug of war" is considered a valid LMW, since it has a definition that could not be inferred from the combination of meanings *tug + of + war* alone. It also undergoes pluralization as a unit – [*tug of war]s* and not *[tug]s of war[s]*. All subterms are valid (bio-)medical terms, and barring syntactic problems like this, should be included in the Lexicon.

Another source of Lexbuild multiwords for more than 20 years, has been in examining expansions of acronyms and abbreviations. All dictionaries include acronyms and abbreviations, of course, and the decision was made to also include their expansions in our Lexicon, unless there is a compelling reason not to. Some types of expansions cannot be Lexbuild records: those that are not a single POS (e.g. "cause of death" and "condition on discharge" are expansions of COD that cannot be LB records); chemical names that are more like formulas than like words (e.g. "1-oleoyl-2-acetyl-sn-glycerol" is an expansion of OAG, but that expansion is not word-like enough to be a Lexbuild record. For adjectives and adverbs, we consider whether the acronym just refers to a sequence of adjectives (or adverbs), as in "circular dichroic" in the adjective record for CD below, or if the acronym refers to an adjective multiword, as with "cardiac denervated" and "choline deficient." We have also declined to make Lexbuild records for names of studies, considering them to be too ephemeral as terms. If those studies have acronyms or abbreviations, the study names can appear as expansions in those records. Whenever non-term expansions occur, they will appear without a corresponding EUI, since they have no corresponding Lexbuild record. Expansions with corresponding Lexbuild records will have that record's EUI appended to the expansion, e.g.:

```
{base=COD
entry=E0453760
        cat=noun
        variants=uncount
        variants=metareg
        acronym_of=cerebroocular dysgenesis|E0453759
        acronym_of=cause of death
        acronym_of=condition on discharge
        acronym_of=chemical oxygen demand|E0453761
        acronym_of=calcium oxalate dihydrate|E0014712
        acronym_of=cystic ovarian disease|E0628408
        acronym_of=chronic oxygen dependency|E0628411
        abbreviation_of=codeine|E0017667
        abbreviation_of=cholesterol oxidase|E0628412
}
```

```
{base=CD
entry=E0516056
```

```
        cat=adj
        variants=inv
        position=pred
        position=attrib(3)
        stative
        acronym_of=circular dichroic
        acronym_of=cardiac denervated|E0747545
        acronym_of=choline deficient|E0747547
}
```

The multiwords we have been adding in the past few years have mostly been based on these prior precedents. Chris Lu makes up lists of high-frequency n-grams, which Destinee, Amanda and Lynn then checked against Essie & other biomedical usage, and make records for only those items that are word-like, a single POS, and show evidence of being terms of art in biomedicine.

How do we decide which **word-like, single-POS items** are terms of art in biomedicine?

1. By checking the Metathesaurus to see if it's included in one or more source vocabularies
2. By checking medical dictionaries for that item, & also for related &/or synonymous terms
3. New acronym and abbreviation expansions are terms
4. Compound nouns are terms; look for characteristic stress pattern, pluralization
5. By analyzing the Hit Patterns array yielded by Essie
6. Terms often come in sets; look for most-used set members in Essie Hit Patterns results

Types 1 through 3 are straightforward. We will expand on types 4 through 6 below.

Compound nouns

Compound nouns are words made of two or more word stems (items that can themselves be independent words), e.g. *ice-cream, heart attack, brother-in-law, stepladder*. Spelling variants often include words separated by space(s), hyphenation, or spelling the compound as a single word:

```
{base=ice cream
spelling_variant=ice-cream
entry=E0033215
        cat=noun
        variants=uncount
        variants=reg
}


{base=heart attack
spelling_variant=heart-attack
entry=E0431577
        cat=noun
        variants=reg
        variants=uncount
annotation=PMID 4124993: history of heart-attack or stroke
}
```

```
{base=brother-in-law
entry=E0014217
        cat=noun
        variants=irreg|brother-in-law|brothers-in-law|
}

{base=stepladder
spelling_variant=step ladder
spelling_variant=step-ladder
entry=E0342486
        cat=noun
        variants=reg
}
```

Though most (count) compound nouns show plural morphology on the last stem, there are some that pluralize the first stem (*brother-in-law/brothers-in-law*) and a few that pluralize both first and last stems (*manservant/menservants*).

Main stress often falls on the first stem, though as length increases, this tends to become less noticeable. Compare, for example, the strong initial stress in *heart surgery* with that of *orthopedic surgery*. In use, stress will also be affected by the sentence and larger linguistic context. So while initial stress will almost always indicate that the sequence in question is a compound noun, there are also compounds in which it is not invariably and clearly present. Linguists do not all agree on where to draw the line between compound nouns and phrases or constructions, though initial stress gives a thumbs-up and certain syntactic tests give a thumbs-down. One syntactic test to keep in mind is whether additional modifiers can be inserted, which phrases do easily, but compounds do not: *black small birds* are not the same as *small blackbirds*. N-grams with initial primary stress and no potential for modifier insertion are the most likely to be compounds, while n-grams without initial primary stress and loose potential for modifier insertion are the least likely to be compounds.

These form two ends of a continuum, and while phrases do not belong in our Lexicon, some of the grey-area n-grams are terms of art, and do arguably merit records. Consider *stem cell*, unambiguously a compound noun. Other clear cases of noun compounds ending in *cell* from Clinical Essie Hit Patterns:

tumor cell
T cell
B cell
HeLa cell
blood cell
muscle cell
cancer cell
Jurkat cell

As the terms get longer, the initial stress is less apparent, yet modification cannot be interposed:

cord blood cell

red blood cell
white blood cell
natural killer cell
umbilical cord blood cell

Although the high-frequency n-grams below have Lexbuild records, they are not compounds:

epithelial cell
endothelial cell
squamous cell
mononuclear cell

High-frequency n-grams that are phrases, not lexical items:

infected cell
inflammatory cell
different cell
single cell

We seem to form new compound nouns with abandon in English (e.g. *shoe bomber*) and our aim is not to enter all such into the Lexicon. However, compound nouns that are biomedical terms of art can be LexMultiwords (LMWs), though LMWs are not limited to compound nouns. We do want to omit descriptive noun phrases, even when they are high-frequency bigrams, unless they are acronym or abbreviation expansions &/or Metathesaurus terms.

Terms of art from Essie Hit Patterns arrays

The Essie search engines developed by Russel Loane and Nick Ide have been invaluable in our analysis of current Medline usage. The most useful Essie tool has been Hit Patterns, which displays in three clear tables the Left Pattern, Right Pattern and Dual Pattern for any given search text. Results are listed in decreasing frequency (of hit count or doc count; we favor sorting by doc count), and selecting any line within a table allows the analyst to drill down into that line's contexts, or to get those sentences or documents.

Left Pattern tables for any single noun X will usually show what types of X Medline writers consider there to be. Consider these n-grams from Historic Medline Essie with *study* as the rightmost member, in decreasing order of doc count frequency; all are >= 300. Asterisked items are not terms:

experimental study
comparative study
clinical study

*contribution to the study
*further study
statistical study
histochemical study
histological study [to be put in]
qualitative study
follow-up study
electrophoretic study [to be put in]
preliminary study
microscopic study
*critical study
biochemical study
radiological study
*clinical and experimental study
pharmacological study
electron microscopic study
chemical study [to be put in]
metabolic study
serological study [to be put in]
microscope study [put in electron microscope study]
laboratory study
*recent study
anatomical study [to be put in]
electroencephalographic study [to be put in]
*experimental and clinical study
immunological study
cytological study

Unsurprisingly, biomedical researchers discern many types of studies, and would likely say that these distinctions are important to science and to their respective subfields. Thus, we argue that terms for types of study belong in our Lexicon. Some of the above could be argued to have compound noun stress, but in general, they are too long to show that pattern clearly. However, they do not allow interposed adjective modification *(*experimental recent study, *radiological recent study, *serological recent study* etc are unacceptable).

The Right Contexts Hit Patterns for the word *study* yield nothing of lexicographic interest, but other search terms have more fruitful Right Contexts patterns. Consider the patterns for *stem cell* in Historic Essie. The most frequent Left Contexts items are *hematopoietic stem cell, neoplastic stem cell, tumor stem cell*, and so on, all of which would be searched in the Metathesaurus and examined in sentence use, to see if they are terms for types of stem cell. The Right Contexts Hit Patterns for *stem cell* show compound nouns and other n-grams beginning with *stem cell*: *stem cell transplantation, stem cell factor, stem cell assay, stem cell factor (SCF), stem cell mobilization, stem cells (PBSC), stem cell support, stem cell rescue, stem cell transplant, stem cell proliferation, stem cells (CFU, stem cells (HSC), stem-cell population, stem cell compartment* and many more. We examine each of these items, by drilling down for further contexts and retrieving sentences using each item. If edited biomedical usage shows us that a given item is being used as a term of art, we make a record for it; if not, we move on to consider other items in the table. As the examples above show, acronym expansions often pop up in these patterns, entirely or in part, and those are either added to existing acronym records, or new acronym records are made which include other expansions shown in Essie searches with the acronym as search text, as well as information from Jablonsky's acronym dictionary & elsewhere.

We also consider Dual Context Patterns. For *stem cell*, the top Dual Contexts item is *hematopoietic stem cell transplantation*, a strong contender as a term, and one that immediately tells us not to accept *hematopoietic stem cell* without convincing evidence that there is such a thing as a hematopoietic stem cell, that is, a stem cell that is hematopoietic. The high frequency of *hematopoietic stem* cell in Essie results could well be because *hematopoietic stem cell transplantation* is a term, i.e. because it is stem cell transplantation that is hematopoietic. Usage indicates that both are terms: the reg plural *hematpoietic stem cells* occurs, and *hematopoietic stem cell transplantation* is a Metathesaurus majorname.

Sets of terms from Essie Hit Patterns

As mentioned above, we can use Left Context Patterns in Essie Hit Patterns to find terms that are "types of X". This holds true whether X is a single word or a compound or other LexMultiWord. In the examples considered above, *stem cell* is a compound noun, and we found evidence of further terms containing *stem cell*, including *hematopoietic stem cell*. More terms might possibly exist, and due diiligence requires that we ask if there is evidence for kinds or types of hematopoietic stem cells

In determining sets of terms, we work from evidence in Medline and other edited texts and avoid speculation. We do not claim that any given set is complete. Rather, some term sets can expand over time as research expands and more types and subtypes of things under study are discerned by scientists. Other sets may contain terms that, while valid, occur so infrequently as not to appear in the text results available to us (We no longer have access to all of Medline due to capacity issues in Essie.), or to appear so infrequently that we are not confident that what little usage we see, represents a consensus. We start with the most frequent items; frequency arguably shows the relative importance of one type of X over another. Among the potential types/subtypes, we do look for relationships, to be as complete as reasonable for the present time. If a disease, body part, protein or whatever, can be termed as X or non-X, we certainly want both types represented in the Lexicon. Lung cancer, for instance, can be termed small-cell lung cancer(/carcinoma) or non-small-cell lung cancer(/carcinoma). The hits for "non-small-cell" outnumber those for "small cell" about 2 to 1, but the mere occurrence of a 'non-X' term points us toward looking for the X term. Again, the facts of usage determine whether we make records for both types of terms. We have not had a hard and fast lower limit of occurrences, but rather try to be reasonable in not wasting time looking for a single decent berry in the brambles, when there is much low-hanging fruit.

Returning to the Hit Patterns in Historic Essie for *hematopoietic stem cell*, we may note several frequently occurring premodifiers to the left of this NP: *allogeneic, human, autologous, pluripotent, murine, primitive, normal, mouse, blood, repopulating, CD34+, derived, bone marrow*. Two of these items look like parts of longer premodifying structures, so we set those aside for now: *blood* and *derived*. The rest all are potential types of hematopoietic stem cell. We determine whether they merit Lexbuild records not only by examining sentences in which the longer NP occurs, but also by considering the array of premodifiers as indicating a potential set of terms. Since we aren't experts in stem cell research, we cannot confidently say exactly what these premodifiers have in common &/or in distinction with one another (though they do seem to refer to sources of the stem cells), nor do we need to. The linguistic evidence shows several subtypes of hematopoietic stem cell. They are not currently in the Lexicon. Whether these should be, is a matter of judgment. Another subtype of hematopoietic stem cell does have a Lexicon record, *totipotent hematopoietic stem cell*, apparently as a result of its being the expansion of the acronym THSC. So, we have *totipotent hematopoietic stem cell* in our Lexicon, but not *pluripotent hematopoietic stem cell*, which occurs far more often. Situations like this have resulted in our adding records for subtypes (sets) of terms, some of which are NP + NP compound nouns and others of which are readily analyzable as adj + NP. As the example of *totipotent hematopoietic stem cell* shows, we already have records for many LMWs with the structure adj + NP, because they are acronym expansions.

Frequency of occurrence and linguistic validity

Not every biomedical term will occur with such high frequency as *stem cell* and its subterms. The Lexicon contains terms for rare diseases, body parts not subject to much research, vectors for diseases cured long ago, uncommon plants, as well as verbs, adjectives and adverbs used to talk about all these. Without them, it would not encompass much of the content of published medical dictionaries and the Metathesaurus, and would not be the gold standard that it is. Therefore, we have applied the principles described above to determine term-hood for newly encountered n-grams that may name rare diseases, rare plants, uncommon disease vectors, and so forth. Their hit counts are lower in Essie searches, but we aim to be consistent in applying the same analytical principles, just on a smaller scale. We do not search out valid but uncommon terms. However, if they come to our attention in the course of making or editing acronym or abbreviation records, or drill-downs of Essie lookups of higher frequency n-grams, we have been making records for them. Once they come to our attention, it has seemed a waste of LSG resources, to shunt aside linguistically valid terms.

The sets of data used during analysis may not fully reflect term frequency. For example, *coronary artery disease* and *computer-aided detection* are both expansions for the acronym *CAD*. In our set of 14,721,393 MEDLINE documents, *computer-aided detection* only occurs 94 times in 64 documents versus *coronary artery disease*, which occurs 51,488 times in 33,802 documents. While within this data set, *computer-aided detection* is infrequent, it is still a valid term. This term could occur more frequently in a different data set, such as a set of radiology reports.

Special Cases

Gene and protein names

We include only those gene and protein names (as uncount nouns) that do not overlap with English words. For example, “ghost”, “desert hedgehog”, and “lava lamp” are common English terms that would not be included in the Lexicon as uncount nouns.

Terms containing slashes, parentheses, and other punctuation marks

We allow special characters such as slashes and parentheses when they are part of a legitimate spelling of a word. The only character we outlaw is a pipe (“|”).

Summary

To summarize, the procedures and principles we have followed in recent years, derive from the basic ones set forth when the Lexicon was originated:

- **Include biomedical vocabulary**, as shown in published dictionaries and the Metathesaurus

- **Reflect biomedical usage**, as shown in Medline abstracts and additional resources not available when the Lexicon was originated, including Google Scholar, and Essie hit patterns showing surrounding linguistic context. Usage informs our coding within each record, as well as pointing us to the existence of new or overlooked biomedical terms.

- **Frequency of occurrence** directs our priorities but has not been a necessary condition for Lexicon term validity, which has been determined on linguistic, rather than statistical principles. Lower frequency terms have been part of the Lexicon since its inception, and while not sought out per se, have been added regularly to the Lexicon if they have been deemed linguistically valid.