Synonym Tagging Guidelines

LSG Linguists' Version

Lynn McCreedy, Fall 2017

2018 Revision: Amanda Payne, Destinee Tormey

We tag Metathesaurus Synonymy Class (sClass) members as to whether the looser notion of synonymy used in each Methesaurus grouping agrees (Y-tags) or does not agree (N-tags) with our much narrower concept of "cognitive synonymy," also called LexSynonymy, explained below. The tagging began in 2015 and should be completed by 2021, after which we only need to tag additions to the Metathesaurus and Lexicon. When tagging, the sClass label is always the concept against which each sClass member is compared, on a binary basis. It is thus important to look up the sClass in Metathesaurus to understand the meaning. We use such a narrow concept of synonymy that any Y-tags within each sClass must also be synonymous with one another. Our basic sources are dictionaries, both general and biomedical. For multiword terms, the entry will be under the head of the construction: for example, for *hypogonadotropic hypogonadism,* look under *hypogonadism.* When terms/sClass concepts cannot be found in dictionaries, Medline usage, Google Scholar usage/articles and Wikipedia are often useful. However, usage should ideally show not only that two terms have exactly the same referent, but that they never have differing referents. When checking dictionaries, be aware that *disease* and *syndrome* are sometimes interchanged, as are the order of names, in terms for diseases or syndromes named after several people. Borderline cases should be N-tagged, to keep as narrow a definition of cognitive synonymy as possible. We do not tag putatively synonymous pairs of genus-species terms; subject area expertise in specific flora/fauna areas would be required for that. Actually, we do tag them: S for "skipped." Items whose Lexicon record has been deleted should be tagged N. Additional Lexical record deletions and edits can also be made by the linguist when duplications or erroneous forms are encountered.

The LSG, focusing our efforts on the Lexicon, has not traditionally concerned itself with the meaning of words or terms, just their variant forms and syntactic behavior. However, the overall aim of the UMLS to be as useful as possible to the NLP community, both within NLM and beyond, has always been the overarching principle. User requests prompted our foray into synonymy in 2015. Users wanted to enhance query results by substituting synonymous terms when initially the query term did not exactly match any Metathesaurus term. A small list of synonyms had been developed in the early 1990s, but remained static. The 2015 effort expanded this through a detailed comparison of Metathesaurus synonym groupings with existing Lexicon terms. It soon became clear that "synonym" itself was a slippery concept. The unabridged Merriam-Webster's offers a number of definitions of *synonym,* ranging from narrow ("one of two or more words of the same language and same grammatical category having the same or nearly the same essential or generic meaning") to broad ("one of two or more words that have one or more senses in common"). The Metathesaurus groupings seemed to contain many items that would pass strict synonymy tests, but also many that seemed to be based on looser concepts of similarity of reference, and broader or narrower meanings than the grouping's concept term. We thus needed to define our own approach to synonyms, which we have termed LexSynonyms.

We take a narrow approach to synonymy, eliminating near-synonyms where one member of the pair has meaning that is broader or narrower than the other. They can thus be said to exhibit meaning-based

commutativity, (x = y) -> (y = x), and **transitivity** ((x = y) and (y = z)) -> (x = z)**.** Commutativity refers to the bidirectionality of the synonymous relationship**.** So, if *Huntington's disease* is a cognitive synonym of *Huntington's chorea,* then *Huntington's chorea* is equally a synonym of *Huntington's disease*. Transitivity extends the equivalence of meaning beyond the pair level. If *Huntington's disease* is a synonym of *Huntington's chorea* and if *chronic progressive hereditary chorea* is also a synonym of *Huntington's disease*, then *chronic progressive hereditary chorea* and *Huntington's chorea* are also synonyms of one another. We have termed these narrowly defined synonyms "cognitive synonyms," but in the tagging context, "synonym" means "cognitive synonym."

A good example of a set of cognitive synonyms is the following:

```
#SYNONYM_CLASS|C0266685|Double monster
128|E0589966|double monster|Y
128|E0589971|twin monster|Y
```

Dorland's defines both *double monster* and *twin monster* as "asymmetrical conjoined twins," so those two terms always and only refer to the same thing. "Asymmetrical conjoined twins" was not included in the Metathesaurus sClass, but had it been, it would also be tagged as (cognitively) synonymous.

Addressing practical methodology, here is a step-by-step demonstration of tagging for a large sClass:

```
#SYNONYM_CLASS|C0269936|Puerperal sepsis
128|E0201783|milk fever|
128|E0214803|puerperal fever|
128|E0217480|puerperal septicemia|
128|E0338998|childbed fever|
128|E0341519|puerperal sepsis|
```

The Y- vs N-tags mark whether each item in the sClass ("#SYNONYM_CLASS") is cognitively synonymous with the sClass term. The first step is to look up the CUI, C0269936, in Metathesaurus. Its semantic type is "Disease or Syndrome." Right off, we can Y-tag the line matching the sClass term because this is a noun for the disease:

```
128|E0341519|puerperal sepsis|Y
```

Dorland's definition of *puerperal sepsis* (a subterm under *sepsis*) yields no exact synonyms within the definition itself; *puerperal septicemia* (under *septicemia*, conveniently on the same page as *sepsis* in Dorland's) does look like it might be strictly synonymous, but this is not confirmed until one follows Dorland's directive to check *puerperal fever*, which it says is synonymous with *puerperal septicemia.* In itself, the synonymy between *puerperal septicemia* and *puerperal fever* is not relevant, unless either of those terms turns out to be synonymous with our sClass term, *puerperal sepsis.* If so, then transitivity would give all of these Y-tags. Checking *puerperal fever,* a subterm under *fever*, yields this: "Called also *childbed f.,* and *puerperal sepsis* or *septicemia"*. This is very helpful; now we can tag all terms except *milk fever* as synonymous with *puerperal sepsis* (and one another). Checking *milk fever* under Dorland's *fever* entry, we see three definitions. This will disqualify *milk fever* from cognitive synonymy with *puerperal sepsis*, because even though one definition seems to refer to something similar –possibly the same—we cannot say that *milk fever* **always and only** means the same thing as *puerperal sepsis.* Here is the tagged sClass:

```
#SYNONYM_CLASS|C0269936|Puerperal sepsis
128|E0201783|milk fever|N
128|E0214803|puerperal fever|Y
128|E0217480|puerperal septicemia|Y
128|E0338998|childbed fever|Y
128|E0341519|puerperal sepsis|Y
```

Common names for genus-species terms present a complication to the "always and only" principle. We mark as "S" those genus-species terms for which we cannot determine synonymy; in practice, nearly all of them. Sometimes an sClass will include common names for genus-species terms, and if that common name is associated both with the sClass term and with a skipped term, we mark that common term "N," since if the two genus-species terms turn out to be nonsynonymous, then the common name is being applied to 2 perhaps similar but nonsynonymous species. Here is an example:

```
#SYNONYM_CLASS|C0324835|Lophocebus albigenia
128|E0535541|Cercocebus albigena|S
128|E0535543|Lophocebus albigena|Y
128|E0535545|gray-cheeked mangabey|N
```

However, it could also turn out that the two genus-species terms are synonymous. In that case, the common name (here, *gray-cheeked mangabey*) is also being applied synonymously. If in the future, the LSG has these sClasses examined by experts on those species, all these tags will be subject to revision.

Some terms can have either a broader or narrower meaning, depending on the specific context of usage. For example, while the term *pons* often refers to the *pons cerebelli* or the *pons Varolii* (which have the same referent in brain structure), *pons* can also refer to any bridgelike anatomical structure. Because *pons* is not in every instance of usage synonymous with *pons cerebelli/pons Varolii, pons* cannot be called cognitively synonymous with either term:

```
pons cerebelli|pons cerebelli|128|pons|128|C0032639|N
pons varolii|pons Varolii|128|pons|128|C0032639|N
pons varolii|pons varolii|128|pons|128|C0032639|N
```

Sometimes, a broader term can have either of two (or more) situationally different meanings, so the broader term is not cognitively synonymous with the narrower one. Take the following pair:

```
coronavirus|corona-virus|128|coronavirus infection|128|C0206750|N
```

One can say one has a *coronavirus* with the same meaning as saying one has a *coronavirus infection*, but *coronavirus* also (some would say primarily) refers to the organism itself, so these two are not cognitively synonymous.

It should be noted that for the present purposes, synonymy extends beyond part of speech boundaries. An adjective can be synonymous with a noun, a verb or an adverb. In this, we are following the practice already established by the Metathesaurus. We both see the adjective *diabetic* as synonymous with the noun *diabetes*.* A Metathesaurus synonym for the noun *malaise* is the verb + adjective construction *feel ill*.

However, because our overall definition of synonymy is narrower, we do not consider some adjectives (and proper nouns used attributively) to be synonymous with nouns with which the Metathesaurus groups them, if their meaning and use indicate a broader or narrower meaning than that of their associated nouns. (Metathesaurus concepts are generally nouns or noun phrases.) For example,

```
#SYNONYM_CLASS|C0001613|Adrenal Cortex

128|E0007478|adrenal cortex|Y
1|E0019186|cortical|N
```

The adjective *cortical* can refer to any of several types of cortex & so does not have synonymy with “adrenal cortex”. The question we should ask ourselves when determining synonymy between an adj/N pair is, “Does this adjective always refer to/have to do with this noun?” For example, *cardiovascular* and *cardiovascular system* are synonymous, because anything that can be described as *cardiovascular* can also be described as having to do with the *cardiovascular system*. *Cardiovascular disease* is *cardiovascular system disease, cardiovascular surgery* is *cardiovascular system surgery, cardiovascular agents* are *cardiovascular system agents* and so on.

To take another example, the relationship of the adjective *deltoid* is not cognitively synonymous to the noun *deltoid muscle*. The Metathesaurus groups *deltoid* in the same synonym class with *deltoid muscle*, and the noun *deltoid* is in fact synonymous with *deltoid muscle*. But is the adjective? According to the unabridged Merriam-Webster’s dictionary, the adjective *deltoid* means ‘shaped like a capital delta’ or ‘constituting or formed like a river delta.’ Biomedical usage in Medline connects its meaning more often with that of *deltoid muscle,* even when that connection is not explicit, e.g. *deltoid ligament, deltoid artery*. Still, we cannot ignore the accepted dictionary definitions. The adjective *deltoid* has those meanings in addition to that connected with *deltoid muscle*, so the adjective and noun are not synonymous.

Concerning proper nouns used attributively, note the N-tag for *Horner* under C0019937 Horner Syndrome while *Huntington* under C0020179 Huntington Disease gets a Y-tag:

```
#SYNONYM_CLASS|C0019937|Horner Syndrome
128|E0000919|bernard's syndrome|Y
128|E0003152|horner syndrome|Y
128|E0204316|horner|N
128|E0238630|bernard syndrome|Y
128|E0238685|horner syndrome|Y
128|E0430341|bernard horner syndrome|Y
128|E0430677|claude bernard-horner syndrome|Y
128|E0501313|oculosympathetic syndrome|Y
```

```
#SYNONYM_CLASS|C0020179|Huntington Disease
128|E0003190|huntington|Y
128|E0003192|huntington's disease|Y
```

```
128|E0016933|chronic progressive hereditary chorea|Y
128|E0421284|huntington's chorea|Y
128|E0431683|huntington chorea|Y
128|E0727407|huntington disease|Y
```

Medline data shows that when used attributively, *Huntington* refers to a Huntington gene, a Huntington protein associated w/Huntington’s disease & chorea, so *Huntington* gets a Y-tag. Contrastingly, *Horner* refers to a Horner muscle & Horner approach, which are not associated with Horner syndrome, so *Horner* gets an N-tag.

Similarly, when an attributive noun can modify a number of head nouns, even nouns referring to the same general topic, the attributive noun cannot be said to be synonymous with all of them, because the transitivity requirement would mean that, for instance, *Alzheimer disease* would be said to be synonymous with *Alzheimer patient*. *Alzheimer* can be said to be synonymous with *Alzheimer disease*, since attributive use of *Alzheimer* always has some connection with the disease. Further synonymy with other MWEs beginning with *Alzheiemer* is not supported by transitivity.

As we see it, the adjective *crispy* is synonymous with the adjective *crisp,* with *crisped* (past tense of the verb *crisp* and by extension, with any other conjugated form of that verb), with the adverb *crisply* and the nouns *crispness* and *crispiness*. Nominalizations of adjectives have the abstract meaning ‘quality or state of being ADJECTIVE’ and can thus always be regarded as synonyms. Likewise, nominalizations of verbs** have the abstract meaning ‘the act, process or instance of VERBing’ and are synonyms of the verbs from which they are derived.

Meaning shifts can and do occur with POS changes in some cases. Consider these examples:

```
#SYNONYM_CLASS|C0003842|Arteries
128|E0010481|arteria|Y
128|E0010531|artery|Y
128|E0694191|arterial|N
1|E0010482|arterial|Y
```

The noun *arterial* refers to roads, not circulatory anatomy, unlike the adjective *arterial*.

```
#SYNONYM_CLASS|C0001774|Agaricales
1024|E0041457|mushroom|N
128|E0041456|mushroom|Y
128|E0221571|champignon|Y
128|E0339151|toadstool|Y
128|E0353801|agaricales|Y
128|E0364407|lycoperdales|Y
128|E0365066|nidulariales|Y
128|E0569047|hymenogastrales|Y
```

The *mushroom* tagged N is the verb (to) *mushroom*, which does not refer to Agaricales.

Synonym tagging also requires us to determine which sense of a term the Metathesaurus Synonym Class Term refers to. Two examples illustrate this:

```
#SYNONYM_CLASS|C4068864|Upbeat
128|E0321064|upbeat|N
1|E0321063|upbeat|Y
#SYNONYM_CLASS|C4068866|See-saw
1024|E0055023|see-saw|N
128|E0055024|see-saw|N
```

In both examples, the Synonym Class term is a "Finding," defined in Metathesaurus documentation as something "discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient." So *upbeat* refers to an emotional/psychological state and *see-saw* refers to relapse or recurrence of a condition. So for *upbeat*, we give a Y-tag to the adjective (E0321063) but must exclude the Lexicon noun record (E0321064) because the noun has other senses, and for *see-saw* we must exclude the noun (E0055023) and verb records (E0055024) because the noun includes other senses.

Whether the synonym class is all nouns, or a mix of nouns with adjective &/or verbs, the basic principle is the same: Are the terms in the group all equivalent terms for the same thing? Always and in every sense? Only those are considered cognitive synonyms, or LexSynonyms.

---

* Although *diabetic* is not listed in the Metathesaurus synonyms for *diabetes mellitus* (and *diabetes* is so listed), interchangeable use of *diabetic* and *diabetes* (or *diabetes mellitus*) is noted in many longer terms:

- diabetic cataract/cataract; diabetes
- diabetic diet/diabetes mellitus diet
- diabetic ketoacidosis/ diabetes mellitus with ketoacidosis
- diabetic nephropathy/nephropathy; diabetes
- diabetic education/diabetes mellitus education/education about diabetes

In their interchangeable use in the expressions above, *diabetic*, *diabetes mellitus,* and *diabetes* are cognitive synonyms that are functioning here as element synonyms. We call them element synonyms because of this very interchangeability in larger multiword expressions. The term element synonym is not directly relevant to synonymy tagging, but see our paper *Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping*.

** There is a major exception to this. Nominalizations of verb particle constructions are considered to be nominalizations of the verb stem in the Lexicon, but these nominalizations are not synonymous with the verb stems. For example, the verb *mock* can occur with the particle *up*, as in these examples from Google Scholar: "the residual interaction is mocked up by a delta interaction"; "The various channel coupling effects for different projectile-target combinations are properly mocked up by the energy dependent potential”. The nominalization of *mock(ed) up* is *mockup* (spelling variants *mock up, mock-up*). This is shown in the Lexicon record for the verb *mock*:

```
{base=mock
entry=E0040636
        cat=verb
        variants=reg
        intran
        tran=np
        tran=pphr(at,np)
        tran=np;part(up)
        nominalization=mockery|noun|E0040637
        nominalization=mockup|noun|E0588063
annotation=Google Scholar: envision a solution, mock it up on paper,
annotation=Google Scholar: mock it up with an inline animation or loop that shows how that would look.
}
```

The nominalization *mockery* is synonymous with the verb *mock*, but the nominalization *mockup* (or its variant forms *mock up, mock-up*) are not. This holds true for all nominalizations which lexicalize the verb particle into the base of the noun. As such, these are exceptions to the general rule that nominalizations are synonymous with their linked verbs/adjectives.