Ambiguity in the UMLS Metathesaurus

2005 Edition

Alan R. Aronson and Sonya E. Shooshan

March 3, 2006

1. Introduction

The UMLS® Metathesaurus® contains a significant amount of ambiguity. For example, the string "cold" (or "Cold" or "COLD") occurs in six distinct concepts with six distinct meanings. The purpose of this report is to examine ambiguity in the 2005AA release of the Metathesaurus in the context of its effect on natural language processing (NLP) applications.

Until the 2004AC release of the UMLS Knowledge Sources, ambiguity was denoted explicitly by appending an ambiguity designator, a number in angle brackets, to the end of an ambiguous string. Thus the ambiguity for "cold" was denoted by 'Cold <1>', 'Cold <2>', 'COLD <3>', 'Cold <4>', 'Cold <5>' and 'Cold <6>'. Now ambiguity is computed by finding concepts with strings that differ only with respect to case.[1]

The degree of Metathesaurus ambiguity has grown over the years and has been especially explosive in 2005. This is probably due both to the presence of SNOMED CT and also to computing

	2000	2001	2002	2003	2004	2005
Strings with an ambiguity designator	9,416	12,840 (+36%)	13,837 (+8%)	16,438 (+19%)	21,295 (+30%)	N/A
Concepts with one or more ambiguity	7,409	9,637 (+30%)	10,328 (+7%)	12,397 (+20%)	16,775 (+35%)	36,113 (+115%)
Concepts with one or more non-suppressible ambiguity	6,501	8,074 (+24%)	8,754 (+8%)	10,416 (+19%)	12,387 (+19%)	33,513 (+171%)
Cases of ambiguity	4,361	5,571 (+28%)	6,014 (+8%)	7,204 (+20%)	10,018 (+39%)	22,218 (+122%)
Cases of non-suppressible ambiguity	4,139	5,311 (+28%)	5,752 (+8%)	6,824 (+19%)	9,521 (+40%)	20,996 (+121%)

Table 1. **Measures of ambiguity in the UMLS Metathesaurus**

ambiguity directly rather than relying on ambiguity designators which may not have captured all

---

1. Note that AMBIGSUI.RRF or AMBIG.SUI cannot be used for this purpose because they do not conflate case.

cases of ambiguity. Table 1 shows several counts that indicate this growth in broad terms. Percentage changes are computed relative to the previous year.

Examining the cases of ambiguity more closely, consider the *degree* of ambiguity, i.e., the number of ways a string is ambiguous or, equivalently, the number of concepts in which it (or one of its case variants) occurs.[1] For example "other" has degree 54 in 2005. Table 2 contains the distribu-

Degree of ambig.	2001 cases	2002 cases	2003 cases	2004 cases	2005 cases
56			1		
55	1	1(0%)			1
54				1	1(0%)
23	1	1(0%)	1(0%)	1(0%)	1(0%)
18	1	2(+100%)	2(0%)	2(0%)	1(-100%)
17	59	58(-2%)	58(0%)	58(0%)	2(-97%)
16	2	2(0%)	2(0%)	2(0%)	1(-50%)
12					1
9					6
8	3	3(0%)	3(0%)	3(0%)	10(+233%)
7	2	5(+150%)	6(+20%)	7(+17%)	11(+57%)
6	4	8(+100%)	8(0%)	7(-13%)	24(+243%)
5	22	21(-5%)	24(+14%)	22(-8%)	54(+145%)
4	55	59(+7%)	79(+34%)	76(-4%)	208(+174%)
3	418	472(+13%)	594(+26%)	600(+1%)	1,239(+107%)
2	4,368	4,641(+6%)	5,541(+19%)	7,250(+31%)	20,659(+185%)
1	635	741(+17%)	885(+19%)	1,989(+125%)	
Total	5,571	6,014(+8%)	7,204(+20%)	10,018(+39%)	22,218(+122%)

Table 2. **Metathesaurus ambiguity distribution by degree**

tion of ambiguities in the Metathesaurus according to degree. Note that an ambiguity of degree one is not actually an ambiguity. In 2004 and before, for examples, 'Abbreviations <1>' is not ambiguous since there were no other 'Abbreviations <n>' strings in the Metathesaurus.

Ignoring suppressible synonyms produces the more realistic distribution shown in Table 3. Most of the ambiguity of degree greater than eight has disappeared, and all of that would disappear if appropriate strings were marked as suppressible. For example, "other" now has degree two. One of these cases, 'Other' (C0205394), is legitimate; and the other, 'Other location of complaint'

---

1. The computation of the degree of an ambiguity was corrected in 2002. As a result, there are some differences from previous editions of this report in the counts reported in the tables.

Degree of ambig.	2001 cases	2002 cases	2003 cases	2004 cases	2005 cases
56					
55					
54					
23	1	1(0%)	1(0%)	1(0%)	1(0%)
18	1	1(0%)	1(0%)	1(0%)	1(0%)
17	1	1(0%)			2
16			1	1(0%)	
12					1
9					5
8	3	3(0%)	3(0%)	3(0%)	8(+167%)
7	2	3(+50%)	4(+33%)	2(-50%)	5(+150%)
6	2	2(0%)	1(-50%)	1(0%)	7(+600%)
5	7	10(+43%)	11(+10%)	7(-36%)	31(+343%)
4	37	46(+24%)	44(-4%)	42(-5%)	156(+271%)
3	385	448(+16%)	473(+6%)	416(-12%)	1,000(+140%)
2	4,511	4,878(+8%)	4,935(+1%)	4,309(-13%)	19,779(+359%)
1	361(+3%)	359(-1%)	1,350(+276%)	4,738(+251%)	
Total	**5,311**	**5,752(+8%)**	**6,824(+19%)**	**9,521(+40%)**	**20,996(+121%)**

Table 3. **Metathesaurus ambiguity distribution after removing suppressibles**

(C0220886), should be marked as suppressible. Suppressible synonyms are ignored for the remainder of this report.

Section 2 of this report describes only the most notable cases of ambiguity in the Metathesaurus, i.e., the cases of degree five or more. The bulk of the cases are now reported automatically by the Migration Assistant, a tool developed generally for annotating ambiguity and specifically for the purpose of marking appropriate cases as suppressible. And Section 3 provides some conclusions derived from this study.

2. Higher Degree Metathesaurus Ambiguity

Ambiguous English Metathesaurus strings are described in this section in decreasing order of degree of ambiguity. Only those cases of degree five or more are covered. See Migration Assistant reports for cases of ambiguity of lesser degree.

In all cases, suppressible synonyms are ignored as is done in Table 3. Ambiguous forms for concepts shown in bold should be marked as suppressible. Recommendations for cases which are not

clear are introduced with the word *consider*. Ambiguous forms for concepts shown in italics should be marked as suppressible in MetaMap only.

2.1 "protocols" (degree 23) <no change from last year>

Except for 'Protocols documentation', the remaining cases should be marked as suppressible because they mean something more specific than "protocols". The concepts involved are

1. C0442711: Protocols documentation
2. **C0542547: Protocols: Activities**
3. **C0677556: Protocols: Pre- or Intra- or Post-Procedure**
4. **C0677557: Protocols: Urinary Elimination**
5. **C0677558: Protocols: Tissue Perfusion**
6. **C0677559: Protocols: Tissue Integrity**
7. **C0677560: Protocols: Sensation, Pain and Comfort**
8. **C0677561: Protocols: Self-Concept**
9. **C0677562: Protocols: Self-Care**
10. **C0677563: Protocols: Safety**
11. **C0677564: Protocols: Role Relationship**
12. **C0677565: Protocols: Respiration**
13. **C0677566: Protocols: Physical Regulation**
14. **C0677567: Protocols: Nutrition**
15. **C0677568: Protocols: Metabolism**
16. **C0677569: Protocols: Medications and Blood Products**
17. **C0677570: Protocols: Immunology**
18. **C0677571: Protocols: Health Behavior**
19. **C0677572: Protocols: Fluid and Electrolyte**
20. **C0677573: Protocols: Coping**
21. **C0677574: Protocols: Cognition**
22. **C0677575: Protocols: Circulation**
23. **C0677576: Protocols: Bowel Elimination**

2.2 "patient education plans" (degree 18) <no change from last year>

All eighteen cases should be suppressed because they are specific kinds of "patient education plans". Their concepts are

1. **C0549081: Patient Education Plans: Activities**
2. **C0549082: Patient Education Plans: Bowel Elimination**
3. **C0549083: Patient Education Plans: Circulation**
4. **C0549084: Patient Education Plans: Coping**
5. **C0549085: Patient Education Plans: Health Behavior**
6. **C0549086: Patient Education Plans: Immunology**
7. **C0549087: Patient Education Plans: Medications and Blood Products**
8. **C0549088: Patient Education Plans: Metabolism**
9. **C0549089: Patient Education Plans: Nutrition**
10. **C0549090: Patient Education Plans: Physical Regulation**
11. **C0549091: Patient Education Plans: Respiration**
12. **C0549092: Patient Education Plans: Role Relationship**
13. **C0549093: Patient Education Plans: Safety**

14. **C0549094: Patient Education Plans: Self-Care**
15. **C0549095: Patient Education Plans: Sensation, Pain and Comfort**
16. **C0549096: Patient Education Plans: Tissue Integrity**
17. **C0549097: Patient Education Plans: Urinary Elimination**
18. **C0549098: Patient Education Plans: Pre- or Intra- or Post-Procedure**

2.3 "assessment" (degree 17)

Except for 'Assessment procedure', the remaining sixteen cases should be suppressed because they are specific kinds of "assessment". The concepts involved in this ambiguity are

1. **C0028708: Nutrition Assessment**
2. **C0542573: Assessment: Bowel Elimination**
3. **C0549068: Assessment: Circulation**
4. **C0549070: Assessment: Coping**
5. **C0549071: Assessment: Fluid and Electrolytes**
6. **C0549072: Assessment: Health Behavior**
7. **C0549073: Assessment: Medications and Blood Products**
8. **C0549074: Assessment: Metabolism**
9. **C0549075: Assessment: Respiration**
10. **C0549076: Assessment: Safety**
11. **C0549077: Assessment: Self-Care**
12. **C0549078: Assessment: Sensation, Pain and Comfort**
13. **C0549079: Assessment: Urinary Elimination**
14. **C0549080: Assessment: Pre- or Intra- or Post-Procedure**
15. **C0679207: Knowledge acquisition using a method of assessment**
16. **C0870300: Assessment: Cognition**
17. C1261322: Assessment procedure

2.4 "ec 2.7.1.112" (degree 17)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0033681: Protein Tyrosine Kinase*
2. *C0065344: Lymphocyte Specific Protein Tyrosine Kinase p56(lck)*
3. *C0109317: EphB2 Receptor*
4. *C0117718: fibroblast growth factor receptor 3*
5. *C0138965: protein-tyrosine kinase c-src*
6. *C0169658: Janus kinase 1*
7. *C0169661: Janus kinase 2*
8. *C0290067: Platelet-Derived Growth Factor alpha Receptor*
9. *C0290068: Platelet-Derived Growth Factor beta Receptor*
10. *C0907648: Ephrin Receptor EphB1*
11. *C0915156: Ephrin Receptor EphA8*
12. *C1333408: EPHA4 protein, human*
13. *C1333409: EPHB3 protein, human*
14. *C1333410: EPHA2 protein, human*
15. *C1334392: LTK protein, human*
16. *C1370509: EPHA1 protein, human*

17. *C1504624: KDR protein, human*

2.5 "ec 2.7.1.-" (degree 12)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0072402: Protein-Serine-Threonine Kinases*
2. *C0108855: cyclin-dependent kinase 2*
3. *C0217310: G Protein-Coupled Receptor Kinase 5*
4. *C0258733: G-protein-coupled receptor kinase 6*
5. *C0645561: CDC7 Cell Division Cycle 7*
6. *C0659150: CHEK1 protein, human*
7. *C0673406: GPRK7 protein, human*
8. *C1333180: Cyclin-Dependent Kinase 10*
9. *C1333735: GPRK2L protein, human*
10. *C1333738: G Protein-Coupled Receptor Kinase Family*
11. *C1337052: PAK6 protein, human*
12. *C1447440: CDK3 protein, human*

2.6 "cam" (degree 9)

Suppress ambiguous form(s) (MetaMap only) because they are either abbreviatory or, in the case of 'CAM brand of Ephedrine Hydrochloride', a brand name. The concepts involved are

1. *C0007578: Cell Adhesion Molecules*
2. *C0054551: cyclophosphamide/doxorubicin/methotrexate*
3. *C0178551: chorioallantoic membrane*
4. *C0678112: CAM brand of Ephedrine Hydrochloride*
5. C0713465: Cam, topical lotion
6. *C1148475: Complementary and alternative medicine*
7. C1366433: CCM1 gene
8. C1366910: Calmodulin 1
9. C1366911: Cerebral Cavernous Malformations 1

2.7 "cat" (degree 9)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. C0007450: Felis catus
2. *C0008169: Chloramphenicol O-Acetyltransferase*
3. *C0040405: X-Ray Computed Tomography*
4. *C0280589: cytarabine/thioguanine*
5. C0325089: Family Felidae
6. C0325090: Felis silvestris
7. C0524517: Genus Felis
8. C1270185: Subfamily Felinae
9. *C1366498: Chloramphenicol Acetyl Transferase Gene*

2.8 "ec 1.14.14.1" (degree 9)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0003805: Aromatase*
2. *C0010763: Cytochrome P-450 CYP2E1*
3. *C0057223: Cytochrome P-450 CYP2D6*
4. *C0059735: Cytochrome P-450 CYP1A1*
5. *C0207509: Cytochrome P-450 CYP1A2*
6. *C0608437: CYP2C9 protein, human*
7. *C0960580: CYP2C19 protein, human*
8. *C1142644: Cytochrome P450 3A4*
9. *C1321893: cytochrome P-450 CYP2C18*

2.9 "ec 3.4.22.-" (degree 9)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0291573: caspase-3*
2. *C0537969: caspase-7*
3. *C0538091: caspase-6*
4. *C0758959: caspase-10*
5. *C0769345: caspase-5*
6. *C0910167: caspase-9*
7. *C0912403: cathepsin Z*
8. *C1437927: CASP4 protein, human*
9. *C1453171: CASP2 protein, human*

2.10 "microsomal monooxygenase" (degree 9)

All nine cases should be suppressed because they are specific kinds of "microsomal monooxygenases". Their concepts are

1. *C0003805: Aromatase*
2. *C0003927: Aryl Hydrocarbon Hydroxylases*
3. *C0010763: Cytochrome P-450 CYP2E1*
4. *C0057223: Cytochrome P-450 CYP2D6*
5. *C0059735: Cytochrome P-450 CYP1A1*
6. *C0207509: Cytochrome P-450 CYP1A2*
7. *C0608437: CYP2C9 protein, human*
8. *C0960580: CYP2C19 protein, human*
9. *C1321893: cytochrome P-450 CYP2C18*

2.11 "cde genotype" (degree 8) <no change from last year>

The eight strings "CDE genotype", "CDe genotype", …, "cde genotype" differ only in the case of "c", "d" and "e". They seem to be legitimate concepts from the Read Codes. In a case sensitive environment, there is no ambiguity. But when case is ignored, as in most NLP applications, the

ambiguity exists. Nevertheless, the ambiguous forms are correctly not suppressed. The concepts involved are

1. C0427651: CDE genotype
2. C0427652: CDe genotype
3. C0427653: CdE genotype
4. C0427654: Cde genotype
5. C0427655: cDE genotype
6. C0427656: cDe genotype
7. C0427657: cdE genotype
1. C0427658: cde genotype

2.12 "cde haplotype" (degree 8)

This case of ambiguity is entirely analogous to the "cde genotype" case. The concepts involved are

1. C1319020: cde haplotype finding
2. C1319021: Cde haplotype finding
3. C1319022: cdE haplotype finding
4. C1319023: CDe haplotype finding
5. C1319024: cDE haplotype finding
6. C1319025: cDe haplotype finding
7. C1319026: CdE haplotype finding
8. C1319027: CDE haplotype finding

2.13 "cde haplotype (finding)" (degree 8)

This case of ambiguity is also entirely analogous to the "cde haplotype" case with the same concepts.

2.14 "cde haplotype finding" (degree 8)

This case of ambiguity is entirely analogous to the "cde haplotype" case and involves the same concepts.

2.15 "driver injured in collision with other and unspecified motor vehicles in nontraffic accident" (degree 8) <no change from last year>

These strings occur in the Read Codes and ICD-10 with some variation in the wording. Except for the last concept below, they should be suppressed because they involve a more specific kind of driver. The concepts involved are

1. **C0476822: Driver pedal cycle injured in collision with other and unspecified motor vehicles in nontraffic accident**
2. **C0476905: Motorcycle driver injured in collision with other and unspecified motor vehicles in nontraffic accident**
3. **C0476973: Car driver injured in collision with other and unspecified motor vehicles in nontraffic accident**

4. **C0477136: Bus driver injured in collision with other and unspecified motor vehicles in nontraffic accident**
5. **C0496239: Driver of three-wheeled motor vehicle injured in collision with other and unspecified motor vehicles in nontraffic accident**
6. **C0496408: Pick-up truck or van driver injured in collision with other and unspecified motor vehicles in nontraffic accident**
7. **C0496481: Heavy transport vehicle driver injured in collision with other and unspecified motor vehicles in nontraffic accident**
8. C0596026: Driver injured in collision with other and unspecified motor vehicles in nontraffic accident

2.16 "driver injured in collision with other and unspecified motor vehicles in traffic accident" (degree 8) <no change from last year>

This case is analogous to the previous one (section 2.15); all but the last should be suppressed. The concepts involved are

1. **C0476826: Pedal cycle driver injured in collision with other and unspecified motor vehicles in traffic accident**
2. **C0476909: Motor cycle driver injured in collision with other and unspecified motor vehicles in traffic accident**
3. **C0476977: Car driver injured in collision with other and unspecified motor vehicles in traffic accident**
4. **C0477140: Bus driver injured in collision with other and unspecified motor vehicles in traffic accident**
5. **C0496241: Three-wheeled motor vehicle driver injured in collision with other and unspecified motor vehicles in traffic accident**
6. **C0496410: Pick-up truck or van driver injured in collision with other and unspecified motor vehicles in traffic accident**
7. **C0496483: Heavy transport driver injured in collision with other and unspecified motor vehicles in traffic accident**
8. C0596027: Driver injured in collision with other and unspecified motor vehicles in traffic accident

2.17 "flavoprotein-linked monooxygenase" (degree 8)

This case is analogous to "microsomal monoxygenase" (section 2.10). All nine cases should be suppressed because they are specific kinds of "flavoprotein-linked monooxygenases". Their concepts are

1. **C0003805: Aromatase**
2. **C0010763: Cytochrome P-450 CYP2E1**
3. **C0057223: Cytochrome P-450 CYP2D6**
4. **C0059735: Cytochrome P-450 CYP1A1**
5. **C0207509: Cytochrome P-450 CYP1A2**
6. **C0608437: CYP2C9 protein, human**
7. **C0960580: CYP2C19 protein, human**
8. **C1321893: cytochrome P-450 CYP2C18**

2.18 "mac" (degree 8)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0009545: Complement Membrane Attack Complex*
2. *C0026916: Mycobacterium avium-intracellulare Infection*
3. *C0065465: cyclophosphamide/dactinomycin/methotrexate*
4. *C0083360: chlorambucil/dactinomycin/methotrexate*
5. *C0279190: cyclophosphamide/doxorubicin/mitomycin*
6. *C0451273: MacAndrew Alcoholism Scale*
7. C0453947: Raincoat
8. *C1167383: membrane attack complex location*

2.19 "allergy" (degree 7)

Suppress 'Allergy brand of chlorpheniramine' (MetaMap only) because it is a brand name. Suppress 'Allergy Specialty' because the speciality is not the same as the disease. Suppress 'Allergy Treatment' because the treatment for a disease is not the same as the disease. The concepts involved are

1. **C0002111: Allergy Specialty**
2. C0020517: Hypersensitivity
3. C0020522: Delayed Hypersensitivity
4. C0020523: Immediate hypersensitivity
5. *C0718383: Allergy brand of chlorpheniramine*
6. **C1262684: Allergy treatment**
7. C1314973: Response to antigens

2.20 "ec 2.7.1.37" (degree 7)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0033640: PROTEIN KINASE*
2. *C0259367: PCTAIRE Protein Kinase 1*
3. *C0294209: LIM Domain Kinase 1*
4. *C0541150: 3-Phosphoinositide Dependent Protein Kinase-1*
5. *C1314894: Col4A3 protein, human*
6. *C1332856: Casein Kinase 2, Alpha 1 Polypeptide*
7. *C1447968: ACVR1 protein, human*

2.21 "mif" (degree 7)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0024429: Migration Inhibitory Factor*
2. C0054504: Calgranulin A
3. C0054505: Calgranulin B
4. C0312359: phenylpyruvate tautomerase

5. C1334507: MIF gene
6. C1335798: S100A9 gene
7. C1366582: S100A8 gene

2.22 "passenger injured in collision with other and unspecified motor vehicles in nontraffic accident" (degree 7) <no change from last year>

This case is analogous to the one in section 2.15 except that all strings should be suppressed. The concepts involved are

1. **C0476823: Pedal cycle passenger injured in collision with other and unspecified motor vehicles in nontraffic accident**
2. **C0476906: Motor cycle passenger injured in collision with other and unspecified motor vehicles in nontraffic accident**
3. **C0496240: Three-wheeled motor vehicle passenger injured in collision with other and unspecified motor vehicles in nontraffic accident**
4. **C0496324: Passenger of car injured in collision with other and unspecified motor vehicles in nontraffic accident**
5. **C0496409: Pick-up truck or van passenger injured in collision with other and unspecified motor vehicles in nontraffic accident**
6. **C0496482: Heavy transport vehicle passenger injured in collision with other and unspecified motor vehicles in nontraffic accident**
7. **C0496486: Bus passenger injured in collision with other and unspecified motor vehicles in nontraffic accident**

2.23 "passenger injured in collision with other and unspecified motor vehicles in traffic accident" (degree 7) <no change from last year>

This case is analogous to the previous one; all ambiguous strings except for the third sense should be suppressed. The concepts involved are

1. **C0476827: Pedal cycle passenger injured in collision with other and unspecified motor vehicles in traffic accident**
2. **C0476910: Motor cycle passenger injured in collision with other and unspecified motor vehicles in traffic accident**
3. C0496242: Passenger injured in collision with other and unspecified motor vehicles in traffic accident
4. **C0496326: Passenger of car injured in collision with other and unspecified motor vehicles in traffic accident**
5. **C0496411: Passenger of pick-up truck or van vehicle injured in collision with other and unspecified motor vehicles in traffic accident**
6. **C0496484: Passenger of heavy transport vehicle vehicle injured in collision with other and unspecified motor vehicles in traffic accident**
7. **C0496488: Passenger of bus injured in collision with other and unspecified motor vehicles in traffic accident**

2.24 "cold" (degree 6) <no change from last year>

The concepts associated with the six senses of "cold" are

1. C0009264: cold temperature
2. C0009443: Common Cold
3. ***C0010412: Cold Therapy*** [consider suppressing]
4. *C0024117: Chronic Obstructive Airway Disease (which has string 'Chronic Obstructive Lung Disease')*
5. C0234192: Cold Sensation
6. *C0719425: Cold brand of chlorpheniramine-phenylpropanolamine*

Senses 3, 4 and 6 are currently suppressed by MetaMap for mainly practical reasons. The only sense that might be marked as suppressible in the Metathesaurus is sense 4, 'Cold Therapy', because it is extremely rare to find the word "cold" by itself meaning 'Cold Therapy'.

2.25 "depression" (degree 6)

Suppress 'Cancer patients and suicide and depression' because it is more specific than "depression". The concepts involved are

1. C0011570: Mental Depression
2. C0011581: Depressive disorder
3. C0344315: Depressed mood
4. C0349217: Depressive episode, unspecified
5. C0460137: Depression motion
6. **C0812393: Cancer patients and suicide and depression**

2.26 "fat" (degree 6)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. C0015677: Fatty acid glycerol esters
2. C0117298: FAT protein
3. *C0279453: doxorubicin/fluorouracil/triazinate*
4. C0424612: Obese build
5. C0812278: FAT gene
6. C1366645: CD36 gene

2.27 "ice" (degree 6)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. C0020746: Ice
2. C0025611: Methamphetamine
3. *C0249492: cytarabine/etoposide/idarubicin*
4. *C0280697: carboplatin/etoposide/ifosfamide*
5. C0534519: Caspase-1
6. C1366479: CASP1 gene

2.28 "mole" (degree 6)

Consider suppressing 'Melanocytic nevus' because it is more specific than "mole". The concepts involved are

1. C0020217: Hydatidiform Mole
2. C0026386: Mole the mammal
3. C0027960: Nevus
4. *C0027962: Melanocytic nevus* [consider suppressing]
5. C0324740: Talpidae
6. C0439189: Mole, unit of measurement

2.29 "premolar tooth" (degree 6)

All strings except for the last one are more specific than "premolar tooth" and should be suppressed. The concepts involved are

1. **C0005373: Bicuspid**
2. **C0447259: Entire maxillary right first premolar tooth**
3. **C0447265: Permanent upper left second premolar tooth**
4. **C0447266: Entire maxillary left first premolar tooth**
5. **C0447299: Entire mandibular right second premolar tooth**
6. C1281668: Entire premolar tooth

2.30 "xenobiotic monooxygenase" (degree 6)

All strings except for the last one are more specific than "xenobiotic monooxygenase" and should be suppressed. The concepts involved are

1. **C0010763: Cytochrome P-450 CYP2E1**
2. **C0057223: Cytochrome P-450 CYP2D6**
3. **C0207509: Cytochrome P-450 CYP1A2**
4. **C0608437: CYP2C9 protein, human**
5. **C0960580: CYP2C19 protein, human**
6. C1412027: Xenobiotic Monooxygenases

2.31 "[so] premolar tooth" (degree 5) <no change from last year>

Like "premolar tooth", all ambiguous strings are more specific and should be suppressed. The concepts involved are

1. **C0005373: Bicuspid**
2. **C0447259: Entire maxillary right first premolar tooth**
3. **C0447265: Permanent upper left second premolar tooth**
4. **C0447266: Entire maxillary left first premolar tooth**
5. **C0447299: Entire mandibular right second premolar tooth**

2.32 "adh" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. Consider suppressing 'Antidiuretic hormome measurement' because the measurement of a substance and that substance are not the same thing. The concepts involved are

1. *C0001942: Alcohol dehydrogenase*
2. C0042413: Vasopressins
3. **C0201849: Antidiuretic hormone measurement** [consider suppressing]
4. *C1332347: Atypical Ductal Breast Hyperplasia*
5. C1366535: AVP gene

2.33 "adjustment" (degree 5)

'Adjustment - classification term' should be reviewed, and is suppressed for MetaMap. The concepts involved are

1. C0001537: Clinical Adjustment
2. C0376209: Individual Adjustment
3. C0456081: Adjustment Action
4. **C0678219: Adjustment - classification term** [needs review]
5. C0683269: Psychological adjustment

2.34 "alp" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. Consider suppressing 'Alkaline phosphatase measurement' because the measurement of a substance and that substance are not the same thing. The concepts involved are

1. *C0102159: alizarinprimeveroside*
2. **C0201850: Alkaline phosphatase measurement** [consider suppressing]
3. C0663932: SLPI protein, human
4. C1366565: SLPI gene
5. C1366566: CCL27 gene

2.35 "apc" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0003315: Antigen-Presenting Cells*
2. *C0033036: Atrial Premature Complexes*
3. C0162832: APC gene
4. C0753723: Adenomatous Polyposis Coli Protein
5. *C1176286: anaphase-promoting complex location*

2.36 "ar" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. C0051755: Amphiregulin

2. *C0332284: Arising in*
3. C1332109: AREG gene
4. C1367578: AR gene
5. C1368501: androgen receptor protein

2.37 "bar" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0001643: beta-2 Adrenergic Receptors*
2. C0441233: External fixator bar
3. C0687760: Taverns
4. C0993613: Bar form
5. C1367657: ADRB2 gene

2.38 "buttercup" (degree 5)

This seems to be legitimate ambiguity.

1. C0330264: Ranunculus
2. C0697205: Ranunculus bulbosus
3. C0939924: Ranunculus bulbosus, Homeopathic preparation
4. C1176287: Cimicifuga racemosa, flower essences
5. C1176288: Ranunculus acris, flower essences

2.39 "cast" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. C0179686: Orthopedic Cast
2. C0302143: Casts body substance
3. *C0687709: Children of Alcoholics Screening Test*
4. C1366481: ASE-1 Gene
5. C1366482: CD3-Epsilon-Associated Protein

2.40 "cd" (degree 5) <no change from last year>

This is legitimate ambiguity (although it is not clear why 'CP protocol' is related to 'CD'). However, the ambiguous forms are suppressed in MetaMap because they are abbreviations. The concepts involved are

1. C0006632: Cadmium
2. C0056447: CP protocol
3. C0079141: Compact discs
4. C0332140: Diagnosis, clinical
5. C0700300: candela

2.41 "cdr" (degree 5)

This seems to be legitimate ambiguity.

1. C0011485: Deoxycytidine
2. C0055351: Chlordecone reductase
3. C1332741: CDR1 gene
4. C1366637: CBFA2T1 gene
5. C1367770: Cerebellar-Degeneration-Related Antigen 1

2.42 "cf" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0009541: Complement Fixation Tests*
2. *C0010674: Cystic Fibrosis*
3. *C0055160: cyclophosphamide/fluorouracil*
4. *C0280055: cisplatin/fluorouracil*
5. C0731033: Leucovorin Calcium

2.43 "cl" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. Consider suppressing 'Chloride measurement' because the measurement of a substance and that substance are not the same thing. The concepts involved are

1. C0010575: Cycloleucine
2. ***C0201952: Chloride measurement*** [consider suppressing]
3. *C0338272: cyclophosphamide/losoxantrone*
4. *C0475212: centiliter*
5. *C0596019: Chloride Ion*

2.44 "dandelion" (degree 5)

This seems to be legitimate ambiguity.

1. C0440015: Dandelion Extract
2. C0877851: Taraxacum officinale
3. C0939896: Taraxacum officinale, Homeopathic preparation
4. C1093500: Taraxacum
5. C1256220: Taraxacum officinale, flower essence

2.45 "dressing" (degree 5)

This seems to be legitimate ambiguity.

1. C0013119: Sterile coverings
2. C0152053: Clothing assistance
3. C0278286: Dressing of skin or wound
4. C0518459: Dressing self-care
5. C1305428: Dressing- activity of daily living

2.46 "erk" (degree 5)

This seems to be legitimate ambiguity.

1. C0109317: EphB2 Receptor
2. C0752312: Mitogen-Activated Protein Kinases
3. C1333340: EPHB2 gene
4. C1366882: MAPK1 gene
5. C1370600: MAP Kinase Gene

2.47 "fundus" (degree 5)

This seems to be legitimate ambiguity.

1. C0016823: Structure of fundus of eye
2. C0017129: gastric fundus
3. C0227254: Fundus of abomasum
4. C0227817: fundus uteri
5. C0740422: Fundus

2.48 "imp" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. *C0021533: Inosine Monophosphate*
2. *C0280103: carboplatin/ifosfamide*
3. *C1150194: mitochondrial inner membrane peptidase activity*
4. *C1264692: Impression AND/OR interpretation of study*
5. *C1328256: mitochondrial inner membrane peptidase complex location*

2.49 "juniper" (degree 5)

This seems to be legitimate ambiguity.

1. C0022431: Juniper extract
2. C0330152: Juniperus
3. C1148550: Juniper
4. C1443050: Juniper antigen
5. C1443051: Juniperus communis, homeopathic preparation

2.50 "lice" (degree 5)

This seems to be legitimate ambiguity.

1. C0023635: Lice
2. C0030756: Lice Infestations
3. C0030757: Pediculus capitis infestation
4. C0030758: Body louse infestation
5. C1058378: Pediculidae

2.51 "mace" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory or a brand name. The concepts involved are

1. *C0280578: cytarabine/etoposide/methotrexate*
2. C0349381: Mace Spice
3. *C0813802: Mace brand of omega-Chloroacetophenone*
4. C0949745: Myristica fragrans
5. C1446581: Mace antigen

2.52 "malt" (degree 5)

Suppress ambiguous form(s) (MetaMap only) because they are abbreviatory. The concepts involved are

1. C0024651: Malt Grain
2. *C0242647: mucosa-associated lymphoid tissue lymphoma*
3. *C0451327: Munich alcoholism test*
4. *C0599921: mucosa-associated lymphoid tissue*
5. C1442914: Malt antigen

2.53 "nutrition" (degree 5)

Consider suppressing 'Nutrition aspect of cancer' because it is more specific than "nutrition". The concepts involved are

1. C0028707: Science of nutrition
2. C0518896: Nutrition outcomes
3. C0600072: Feeding and dietary regimes
4. C1442959: Nutrition function
5. ***C1443053: Nutrition aspect of cancer*** [consider suppressing]

2.54 "olive" (degree 5)

This seems to be legitimate ambiguity.

1. C0228539: Medulla oblongata olive
2. C0262254: inferior olive
3. C0349382: Olives - dietary
4. C1116536: Olive, Bach Flower Essence
5. C1365464: Olive extract

2.55 "pi3k" (degree 5)

This seems to be legitimate ambiguity.

1. C0044602: 1-Phosphatidylinositol 3-Kinase
2. C0285761: Phosphoinositide-3-Kinase, Catalytic, Gamma Polypeptide
3. C1335214: PIK3CG gene
4. C1368105: Phosphoinositide-3-Kinase, Catalytic, Beta Polypeptide
5. C1368106: Phosphoinositide-3-Kinase, Catalytic, Alpha Polypeptide

2.56 "plantain" (degree 5)

This seems to be legitimate ambiguity.

1. C0032094: Plantago major
2. C0946846: Plantago
3. C1039591: Musa x paradisiaca
4. C1235662: Plantain
5. C1365964: Common plantain preparation

2.57 "pressure" (degree 5)

This seems to be legitimate ambiguity.

1. C0033095: Pressure- physical agent
2. C0234222: Baresthesia
3. C0392677: Pressure
4. C0460139: Pressure - action
5. C1306345: Pressure application

2.58 "ptc" (degree 5)

This seems to be legitimate ambiguity.

1. C0015491: Factor IX
2. C0694887: PTCH gene
3. C0694890: RET gene
4. C1366464: F9 gene
5. C1366467: D10S170 gene

2.59 "screening" (degree 5)

Consider suppressing 'Screening for cancer' because it is more specific than "screening". The concepts involved are

The concepts involved are

1. *C0199230: Screening for cancer* [consider suppressing]
2. C0220908: Screening procedure
3. C0220909: Aspects of disease screening
4. C1305399: Screening - procedure intent
5. C1409616: Special screening finding

2.60 "silver" (degree 5)

Consider suppressing 'Silver measurement' since the measurement of a substance and that substance are not the same thing. The concepts involved are

1. C0037125: Silver
2. *C0202473: Silver measurement* [consider suppressing]
3. C1161330: Argentum metallicum, silver, Homeopathic preparation
4. C1305285: Silver color
5. C1318876: Silver stain

2.61 "vitamin" (degree 5)

Except for the fourth and fifth sense, each of these is a specific kind of "vitamin" and should be suppressed. The concepts involved are

1. **C0042839: Vitamin A**
2. **C0042866: Vitamin D**
3. **C0042878: Vitamin K**
4. C0042890: Vitamins
5. C0087096: Tocopherols

3. Conclusions

Some concepts have ambiguous forms which should be marked as suppressible, where *ambiguous forms* means both the string with an ambiguity designator and one or more strings without it. Note that some concepts such as 'Assessment: Cognition' have more than one ambiguous form with ambiguity designator ('Assessment <5>' and 'Assessment <17>' in this case). The analysis in this and previous editions of this report reveals some classes of ambiguity commonly occurring in the Metathesaurus:

- **Contextual (or hierarchical) ambiguity.** This class of false ambiguity is exemplified by the string 'Prostate' for 'Prostatic Diseases'. It normally arises from terms which require context within their vocabulary (in this case, a disease hierarchy) in order to be properly understood. Contextual ambiguities can be classified according to their participants:
  - **Body part/disease ambiguity** exemplified by 'Prostate' and 'Prostatic Diseases'
  - **Body part/procedure ambiguity** exemplified by 'Stomach' and 'Procedures on the stomach'
  - **Pathology/procedure ambiguity** exemplified by 'Pathology' and 'Pathology procedure'
  - **Medical device/procedure ambiguity** exemplified by 'Prosthesis' and 'Prosthesis Implantation'
  - **Substance/therapy ambiguity** exemplified by 'Anthracyclines' and 'prior anthracycline therapy'
  - **Substance/measurement ambiguity** exemplified by 'Thyroid stimulating immunoglobulins' and 'Thyroid stimulating immunoglobulins assay'
- **Generalization ambiguity.** This is also false ambiguity caused by grouping several concepts together using a more general term. For example, 22 concepts including 'Protocols: Activities' and 'Protocols: Pre- or Intra- or Post-Procedure' are generalized to 'Protocols' which does seem to be a legitimate synonym of the concept 'Protocols documentation'.
- **Meta ambiguity.** This new class of ambiguity, represented by strings such as 'Other complications of procedures NEC in ICD10', contain meta information. In this case it is the name of the vocabulary, ICD10 in the example. As opposed to the first class of ambiguity above in which strings such as 'Prostate' meaning 'Prostatic Diseases' do not say enough about themselves, these strings say too much. It is true that the meaning of a string containing 'NEC', 'not elsewhere classified' or like phrase, depends upon its vocabulary, but such information is already available in the MSRO file (where it belongs). It is also true that such strings have different meanings and strictly speaking should be different concepts. But the practical result of such a representational scheme is to introduce an ambiguity that most users do not want or need to

resolve. (It is not even clear that those who might want to resolve the ambiguity can do so with the information available in the Metathesaurus.)

- **Abbreviation ambiguity.** This is another, large class of ambiguity caused by distinct concepts having the same acronyms (or abbreviations). An example from above is that 'Mitral Valve Stenosis', 'Multiple Sclerosis', 'Morphine Sulfate' and 'millisecond' all have abbreviation 'MS' or 'ms'. Although this class represents true ambiguity in a strict sense, it is better to disallow it in many text processing situations, especially those in which authors define the abbreviations they use. Unlike the other classes of ambiguity defined above, we do not recommend that this case be reflected in changes to the Metathesaurus. This kind of ambiguity will be suppressed for MetaMap processing only.

One limitation of this study is that there was no follow-up of ambiguous strings discovered while examining the strings for a given case. For example, the set of all strings for the concepts containing a 'sound measurement <n>' string include the ambiguous strings 'Ultrasonography <n>', 'Echography <n>', 'Ultrasound <n>', 'Echotomography <n>' and 'ultrasound scanning <n>'. If the ambiguous forms for 'sound measurement <n>' are to be marked as suppressible, how does that affect the other ambiguous strings? Each of them is part of a cluster of concepts, possibly different from the one for 'sound measurement <n>'. Although it is probably not necessary to explore the new concept clusters, it is essential to examine the original cluster for additional ambiguous forms to suppress.