Hi Everyone:

We analyzed the results of LUI assignment based on Soma's reports (Merge, Split, and Split_Merge) to monitor the behavior of LuiNorm, enhance LuiNorm and its associated LVG flow components, and update Lexicon data for the next release. In this analysis on 2009 version, as expected, we observe the total numbers of all three cases (merge, split, split-merge) are dramatically decreased compared to last release. The reason is that there is only one software change on the LuiNorm tool. Which is the change of database on Canonical form generation program (from MySql to HSqlDb) to improve the handle ability on Unicode characters in LuiNorm operation. In other words, the results of LuiNorm should be the same between 2008 and 2009 releases except for the input contains:

- Unicode characters
- New words from LEXICON or UMLS - atoms

In our study, we found there is no Unicode character in any of merge, split, and split_merge cases. All these three cases with different LuiNorm results are caused by the different base forms or Canonical forms resulting from new words in UMLS, new words in LEXICON, new inflectional rules in LEXICON, and new spelling variants in LEXICON. These behaviors are expected to happen every year and considered as system enhancement.

Also, we found five merge cases and one split_merge case with same LuiNorm results. Theoretically, terms have same LuiNorm results should have same LUI assignment between releases. Accordingly, we assume there are some software changes on the Lui-Assignment program. Also, we try to guest these changes based on our observation. Please refer to the analysis sections of Merge and split_merge bellowed for details.

In conclusion, luiNorm.2009 behaves very well and no software change requests or lexical records updates are found from on this study. We are pleased with the results and considering the algorithm of LuiNorm is stable. We expect the numbers of all three cases will be in the same magnitude next year. Please refer to NLM internal web page at URL http://lexlx1.nlm.nih.gov/LexSysGroup/Projects/lvg/2009/docs/designDoc/LifeCycle/test/luiAssign/index.html for the details if you are interested.

==========

- **Merge cases:**
  - Summary:
    There are 2618 SUIs with 673 Luis merged into 278 Luis. As expected, these numbers have been dropped between 40% to 70% compared to the previous release (2008). Table 1 shows the percentage distribution of all causes of merge cases. There are 78.42% and 19.78% of merge cases caused by the changes of Canonical forms and base forms, respectively. Please note that we found 5 merges cases (1.80%) with same LuiNorm results for both 2008 and 2009 releases. Please refer the analysis section below for details.

Causes of merge cases	Merge No.	Percentage	Examples
Canonical form (new data in LEXICON & UMLS)	218	78.42%	▪ Composti ▪ Vestibulus ▪ Tetrazole ▪ foetography
Base form (new data in Lexicon)	55	19.78%	▪ cavae ▪ washings
Others, not caused by LuiNorm	5	1.80%	▪ Act <10> ▪ Arm <10> ▪ cancer <9> ▪ conjunctiva <10> ▪ cornea <9> ▪ cornea <10>

Table 1. Percentage distribution of merge causes

- Merge analysis:

  1). By Canonical form (caused by new data in UMLS - atoms or LEXICON):
  We changed the database of canonical form generation program from MySql to HSqlDb to improve the handle ability on Unicode characters in 2009. There is no other software change. In other words, the results of canonical forms should be the same as the previous release except for:
  - o inputs contain non-ASCII (Unicode) characters
  - o inputs contain new data from UMLS – atoms or LEXICON

  In our study, there is no Unicode character in merge, split, and split_merge cases. In a word, all these cases are caused by data change. Below, we illustrate two examples in merge cases for these causes:

  Example 1 - New words in UMLS
  "composti" is a new word in UMLS. It belongs to same canonical class with "compost" and "compostus" in 2009 while it was not in 2008. Accordingly, "composti" is merged with "compostus" into one LUI in 2009.

  Example 2 - new spelling variants in LEXICON
  "tetrazol" is added to the Lexical record (E0205191) as a new spelling variant of "tetrazole" in 2009. Accordingly, "tetrazole" is merged with "tetrazol" into one LUI.

  2). By base forms (new data in Lexicon):
  The results of base forms from Lexical tools mainly depend on the data of LEXICON. The base form of a word might be different from last version if there are new lexical record(s), modified or new inflectional rules, or deleted lexical records associated with this word. These cases are expected to be observed

every year and are considered as an enhancement between releases. Cases in this category have 19.78% (55/278) impact of merge cases in 2009 release. Two examples are illustrated as follows:

Example 1 - New Lexical Records
"cava" is a new record (E0015668) in LEXICON 2009. According to the inflection rules (variants=glreg) in the lexical record, "cava" has base form of "cavae" and merged together to the same LUI.

Example 2 - Modified lexical records with new inflection rules
The lexical record (E0065082) of "washing" is modified by adding a new inflection rule 'variants=reg' in 2009 lexicon. Accordingly, "washings" is uninflected to "washing" and then canonicalized to "wash". Accordingly, "washings" merges with "wash" to have same LUI.

3). There are five new merged cases (six Luis) have same luiNorm results from both LuiNorm 2008 and 2009. They are:

- Act <10>
- Arm <10>
- cancer <9>
- conjunctiva <10>
- cornea <9> and cornea <10>

Theoretically, terms have same LuiNorm results should not have different LUI assignment between releases. Also, we observed all the above merge cases involved ambiguity tags of <9> and <10>. Accordingly, we assume there is a software change on the LUI assignment software on handling ambiguity tags of <9> and <10>.

==========

- **Split:**
  - Summary:
    There are 230 SUIs with 65 LUIs split to 131 LUIs. As expected, these numbers have been dropped about 70% compared to the previous release (2008). Most of these split cases (83.08%, 54/65) are caused by different Canonical forms. As mentioned above in Merge section, this is caused by the data change in UMLS - atoms and LEXICON. In addition, new base forms from new lexical records in Lexicon introduced 16.92% (11/65) on the split cases. Table 2 shows the percentage distribution of each cause. We did not find any unexpected behavior of LuiNorm in the split cases.

Causes	**Split No.**	**Percentage**	**Example**
Canonical form (new data in UMLS & Lexicon)	54	83.08%	▪ Cyrtopodium ▪ Allogeneic

			▪ metamfetamine
Base form (new data in Lexicon)	11	16.92%	▪ posset ▪ microfabrication

Table 2. Percentage distribution of split causes

- Split analysis:
  1). By Canonical form:
  As mentioned in the merge cases, different canonical forms can be caused by the data change in UMLS - atoms or LEXICON if the input does not contains Unicode characters. Bellows, we illustrate three examples of split cases caused by different data change in resulting different Canonical forms:

  Example 1 – New words in LEXICON
  A new record, “Cyrtopodion” E0647568, is added into Lexicon 2009 as a noun. The rule of generating inflectional variant “cyrtopodia” from “cyrtopodion” as a noun in 2008 is not valid in 2009. Thus, “cyrtopodion” splits from “cyrtopodium” to have different LUIs.

  Example 2 – Spelling variant is split into a new record in LEXICON
  “allogeneic” is removed from “E0008162” as a spelling variant of “allogenic” and added as a new record (E0628375) into Lexicon 2009 as an adjective. Thus, “allogeneic” splits from “allogenic” to have different LUIs.

  Example 3 – change in spelling variants:
  The spelling variant, “metamfetamine” of lexical record (E0039911) “methamphetamine” is removed and merge into lexical record (E0530856) “metamphetamine” as its spelling variant in 2009. Accordingly, “metamfetamine” is split from “methamphetamine” to have different LUIs.

  2). By base forms (new data in Lexicon):
  Split cases can be caused by new base forms from new lexical records. For example, “posset” is a new lexical record (E0623055) as verb with regular inflection rule in 2009. Thus, “possetting” is uninflected to “possett” in2009 (by fact) while it was uninflected to “posset” in 2008 (by rules). Accordingly, “possetting” is split from “posseting” to have different LUIs.

==========

- **Split_Merge:**
  - Summary:
    Split merge cases are the cases when some words split first, then merge (with others) again. There are 315 SUIs with 13 LUIs split_merge cases. The total number (13) of split merge cases in 2009 is relatively small of LUI assignment. Table 3 shows the percentage distribution of each cause by LuiNorm flow components. There are 84.62% (11/13) and 7.69% (1/13) caused by new base

forms and new canonical forms, respectively. Please note that there is one split_merge case with same result from LuiNorm. Please refer to the detail analysis below.

Causes	Split_Merge No.	Percentage	Example
Base form (new data in Lexicon)	11	84.62%	▪ ESS ▪ VES ▪ ESSS ▪ PKCS
Canonical form (new data in LEXICON & UMLS)	1	7.69%	▪ VED
Others, not caused by LuiNorm	1	7.69%	▪ stopwords

Table 3. Percentage distribution of split_merge causes

- Split_Merge analysis:
  The cause of this case is the combination of above two (split and merge). Potentially, terms in the split_merge cases might belong to same canonical class. This is the data we use to enhance the algorithm of canonical form and make canonical class covers bigger range (more words). In this analysis (2009), we did not found anything to enhance luiNorm from these 13 cases. However, we found there is one split_merge case with same LuiNorm result. Bellows are the brief discussion on these causes.

  1). Base forms:
  Please refer to the discussion in split or merge cases.

  2). Canonical forms:
  Please refer to the discussion in split or merge cases.

  3). Others:
  We observed stopwords, such as “with”, “NOS”, “in”, “by”, “and”, etc.. are in this one split_merge cases. The LuiNorm results on these stopwords are the same as last year. As discussed in Merge cases above, we assume there is a software change on the Lui-Assignment program to cause these split_merge cases.