Hi Everyone:

We analyzed the results of LUI assignment based on Joanne'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 2008 version, we observed the total numbers of all three cases (merge, split, split-merge) are all decreased compared to last year. In a sense, it means luiNorm is toward to a stable state. No software change request on lexical tools or lexical records update in Lexicon is suggested for the next release (2009) from this analysis. We are pleased with the results and considering the algorithm of luiNorm is stable and closer to optimum.

In 2008 Lexical Tools, there are two major software change associated with luiNorm. They are:

- Bug fixed on Canonical form generation program to include correct version of spelling variants and lexicon records into the same canonical class.
- Enhanced flow components of normalizing non-ASCII characters

First of all, the bug fixed on Canonical form generation program is the major cause on all three cases: 511 out of 886 merge cases; 149 out of 207 split cases; and 15 out of 29 split-merge cases. All changes caused by this fixed should not show up in the next year LUI assignment. Second, the enhanced flow component of –f:q7 and –f:q8 caused less than 7% of merge cases and 0 % on split or split-merge cases. In other words, the enhanced non-ASCII normalization flow components behave as designed and do not have too much impact and cause any noise on LUI assignment. Similar to previous versions, certain small portion of all three cases are caused by data change in Lexicon (new base form and spelling variants). Last, we found few unexpected behavior in the split-merge cases and led us to find and fix a small bug in the LUI assignment program.

In conclusion, luiNorm.2008 behaves very well and no software change requests or lexical records updates are found based on this report. If there is no major software change on the next release of Lexical tools, we expect the change on all three cases will be decreased next year. Please refer to NLM internal web page at URL http://lexlx1.nlm.nih.gov/LexSysGroup/Projects/lvg/2008/docs/designDoc/LifeCycle/test/luiAssign/index.html for the details. Bellows are the detail discussion of 2008 analysis if you are interested:

==

- **Merge cases:**
  - Summary:

    There are 4431 SUIs with 1817 Luis merged into 886 Luis. More than 76% of these merges are due to the change of Canonical forms. In addition, new base forms from new lexical records in Lexicon contribute more than 16% for the merge cases. Only less than 7% (5.98% + 0.56%) of merge cases are caused by the new enhanced non-ASCII normalization flow component. Table 1 shows the percentage distribution of all causes to merge cases. We did not find any unexpected behavior of luiNorm from the merge cases.

Causes of merge cases	**Merge**	**Percentage**	**Examples**

	No.		
Canonical form (bugs fixed & new data)	682	76.98%	▪ `chlorethanol` ▪ `chloroethanol`
Base form (new records in Lexicon)	146	16.48%	▪ `aniseikonias` ▪ `Anomias`
New flow –f:q8 to replace –f:q4	53	5.98%	▪ `Actos®` ▪ `Nexium™`
New flow –f:q7 to replace –f:q:q2	5	0.56%	▪ `No brake space` ▪ `’`

Table 1. Percentage distribution of merge causes

- Merge analysis:

  1). Canonical form (682/886, 76.98%):

  There is a bug fixed of utilizing correct spelling variants and lexical data in the canonical form generation program in 2008. This fix enhanced the luiNorm form by including all spelling variants in the same canonical class and introduced about 75% (511/682) of the merge cases caused by canonical form change. In addition, new lexical records (new EUI) and new spelling variants resulted in different canonical forms and contributed 18.18% and 6.89% of canonical form change, respectively. We run a detail analysis on the causes of the change on canonical forms and the results are shown in table 2.

Causes	Merge No.	Percentage	Example
Bug fixed	511	74.93%	▪ `2 chlorethanol` ▪ `2-chlorethanol` ▪ `2-chloroethanol`
New lexical records (EUI)	124	18.18%	▪ `Disulphiram` ▪ `FLAVINE`
New spelling variants	47	6.89%	▪ `Hypsarrhythmia` ▪ `Skatole`

Table 2. Percentage distribution of detail causes on Canonical forms of merge cases

  2). Base forms (new data in Lexicon):

  The results of base forms from Lexical tools mainly depend on the data of The SPECIALIST LEXICON. The base form of a word might be different from last year if there are new lexical record(s), modified or new inflectional rules, or deleted lexical records associated with this word. Accordingly, this case is expected to be observed every year and is considered as an enhancement between releases. This case has 16.48% (146/886) impact of merge cases in 2008 release.

  For example:

  - New Lexical Records:

  L1713632|L1713632|S1940519|PCOS

"PCO" is a new word (E0591949) in LEXICON 2008. According to the inflection rules in this lexical record, "PCOS" is uninflected to "PCO" and merged with string "PCO" to have same LUI.

- Modified lexical records with new inflection rules:

L6321687|L0003113|S0014244|Anomias
The lexical record of "anomia, E0009197" is modified by adding a new inflection rule 'variants=reg' in 2008 lexicon. Accordingly, "anomias" is uninflected to "anomia" and merges with "anomia" to have same LUI.

3). New enhanced non-ASCII normalization flow components (-f:q7 and –f:q8):
One of the major software upgrade in the Lexical tools 2008 is the normalization on non-ASCII characters. Flow, -f:q7, Unicode core norm, is used to replace –f:q and –f:q2 to strip diacritics and split ligatures. It also enhances normalization by mapping Unicode punctuation and other Unicode characters. For example, right single quotation mark in the "Princess Mary’s Royal Air Force …" is normalized to ASCII character apostrophe.

Flow, -f:q8, strips unknown Unicode or maps known Unicode, is used to replace –f:q4 at the end of luiNorm to ensure the pure ASCII result from luiNorm. For example, registered sign ® and trade mark sign ™ are stripped in 2008 release. In other words, "BeneFin" and "BeneFin™" are all normalized to "benefin" and have same LUI.

===

- **Split:**
  - Summary:
    There are 752 SUIs with 207 LUIs splited to 420 LUIs. Most of these split cases (77.29%, 160/207) are caused by different results of the Canonical form. As mentioned above, there is a bug fixed in the 2008 Canonical form generation program to include all correct spelling variants and lexical records in the same canonical class. This fix caused the major part of Canonical form change in the split cases. In addition, new base forms from new lexical records in Lexicon introduced 22.71% (47/207) on the split cases. The new enhanced non-ASCII flow components, -f:q7 and –f:q8, did not cause any split cases. This means the new enhanced non-ASCII flow components behave properly and do not introduce any unexpected noise. Table 3 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 algorithm & data)	160	77.29%	▪ `miosis`
Base form (new data in Lexicon)	47	22.71%	▪ `Drakeol` ▪ `Drakeols` ▪ `somnography`

Table 3. Percentage distribution of split causes

  - Split analysis:
    1). Canonical form:

As mentioned in the merge cases, there is a bug fixed of utilizing correct spelling variants and lexical data in the canonical form generation program in 2008. This fix enhanced the luiNorm form by including all spelling variants in the same canonical class. This fix contributed most of the split cases caused by canonical form change (149/160, 93.13%). In addition, new lexical records (new EUI) and new spelling variants resulted in different canonical forms and contributed 6.25% and 0.62% of canonical form change, respectively. We run a detail analysis on the causes of the change on canonical forms and the results are shown in table 4.

Causes	**Merge No.**	**Percentage**	**Example**
Bug fixed	149	93.13%	▪ `Meiosis`
New lexical records (EUI)	10	6.25%	▪ `ACLY` ▪ `polyvisoline`
New spelling variants	1	0.62%	▪ `metaupon`

Table 4. Percentage distribution of detail causes on Canonical forms of split cases

- Example 1 – Bug fixed:
  The spelling variants, "miosis" and "myosis", of lexical record E0039359 "meiosis" are removed and split into a different lexical record (E0590267). Accordingly, "miosis" and "myosis" are split from "meiosis" to have different LUIs. These records were correctly used in 2008 Canonical form generation program.

- Example 2 – New Lexical records:
  A new record, "preolivary" E0612013, is added into Lexicon 2008 as an adjective. The rule of generating inflectional variant "preolivaris" from "preolivary" as a noun is no longer valid. Thus, "preolivary" splits from "preolivaris" to have different LUIs.

- Example 3 – New spelling variants:
  The spelling variant, "metopon" of lexical record E0304676 "metaupon" is removed and split into a new lexical record (E0612463) in 2008. Accordingly, "metopon" is split from "metaupon" to have different LUIs.

2). Base forms (new data in Lexicon):

Split cases can be caused by new base forms from new lexical records. Please refer to the discussion in merge cases.

==

- **Split_Merge:**
  - Summary:
    Split merge cases are the case when some words split first, then merge (with others) again. There are 692 SUIs with 29 LUIs split_merge cases. The total number (29) of split merge cases in 2008 is relatively small of LUI assignment. The bug fixed in the canonical form generation program takes more than 55 % (16/29). Different base forms caused by new

lexical records have 27.59% (8/29). There are 5 split merge cases (17.24%) that are not caused by the difference of luiNorm output. As we traced down, we found a wrong flag setting in the LUI assignment program. This bug is corrected to regenerated the LUI assignment reports. This analysis is based on the new updated reports.

Tables 5 and 6 show the percentage distribution of each cause by luiNorm flow components and detail causes on Canonical form change. We did not observe any unexpected behavior of luiNorm in the split_merge cases.

Causes	**Split_Merge No.**	**Percentage**	**Example**
Canonical form (new algorithm & data)	16	55.17%	• `Bufenine` • `Buphenine` • `Buphenin`
Base form (new data in Lexicon)	8	27.59%	• `FAC`
Others, not caused by LuiNorm	5	17.24%	• `AN <9>` • `AN-9`

Table 5. Percentage distribution of split_merge causes

Causes	Split_Merge No.	Percentage	Example
Bug fixed	15	93.75%	• `Bufenine` • `Buphenine` • `Buphenin`
New EUI records	1	6.25%	• `A (E0598104)` • `A (E0598106)`

Table 6. Percentage distribution of detail causes on Canonical forms of split merge cases

- Split_Merge analysis:
  The cause of this case is the combination of above two (split and merge). Potentially, terms in these 29 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) when it makes sense. We did not found anything to enhance luiNorm from these 29 cases in 2008. However, we found there is a software inconsistency on the LUI assignment program due to a wrong flag setting. This program is corrected to regenerate the correct LUI assignment.

  1). Canonical form:
  Please refer to the discussion in split or merge cases.

  2). Base forms (new data in Lexicon):
  Please refer to the discussion in merge cases.

  3). Others:

The analysis program finds the difference on the results of luiNorm and each step (results of its associated flow components) in the luiNorm and tags them. It is tagged as “others” when no difference from flow components is identified. In other words, the luiNorm results of these words are the same. Accordingly, these terms should not have new split merge cases. These cases are considered as software inconsistency issue (leads to a potential bugs in LUI assignment program). All such cases should be manually examined and trace down the cause. In this analysis, we found a wrong flag setting in luiNorm assignment program.