Hi Everyone:

As usual, we analyzed the testing results of luiNorm.2007 based on Brian's report (Merge, Split, and Split_Merge) to monitor the behavior of luiNorm, enhance luiNorm and associated lvg flow components, and update Lexicon data for the 2008 release. In this analysis, no software change request is suggested for the next release. Only few possible duplicated lexical records are found and will be used to purify Lexicon database.

In lvg2007, there are two major software change associated with luiNorm based on the analysis report of 2006 luiAssignment. They are:

- Enhancement on flow component, -f:rs, to remove upper case parenthetical plural forms, such as (S). (ES), and (IES). This software change introduced 850 (73.34%) of merge cases. These 850 merge cases are considered as enhancements on LUI assignment.
- Enhancement on Canonical generation program to include all spelling variants into the same canonical class and select the word with min. length from the same canonical class as the canonical form. This software change introduced 6617 (99.19%) of split cases. These splits resolve the issue of same spelling between words and abbreviations. They are also considered as enhancements on LUI assignment.

There are small portion of merge, split, split_merge cases caused by the data change in Lexicon other than above two causes. These changes are normal results and expected to occur on every new release. In a word, luiNorm.2007 behaves very well and no software change requests are suggested based on this report. Please refer to NLM internal web page at URL `http://lexlx1.nlm.nih.gov/LexSysGroup/Projects/lvg/2008/docs/designDoc/LifeCycle/test/luiAssignment.html` for the details of luiAssignment analysis. Bellows are the detail discussion of 2007 analysis if you are interested:

==========

- **Merge:**
  - Summary:
    There are 3538 SUIs with 2330 Luis merged into 1159 Luis. Most of these merges are due to the enhancement on the lvg flow component of removing upper case parenthetical plural form (-f:rs). This is a feature enhancement we implmented based on the report of last year (2006) luiAssignment. In addition, new base forms from new data in Lexicon and enhancement of Canonical form generation program are two other factors for the merge cases. Table 1 shows the percentage distribution of each cause. We did not find any unexpected behavior of luiNorm from the merge cases.

Causes	Merge No.	Percentage	Example
Enhancement on the lvg flow –f:rs	850	73.34%	`TRIM NAIL(S)`
Base form (new data in Lexicon)	125	10.78%	`Apis`
Canonical form (new algorithm & data)	184	15.88%	`Synerone, synera`

Table 1. Percentage distribution of merge causes

- Merge analysis:

1). Lvg Flow component, –f:rs

Based on the report of 2006 luiAssignment, we enhanced the lvg.2007 flow of removing parenthetical plural forms, -f:rs, to remove the upper case parenthetical plural forms, such as (S), (ES), and (IES). This software change takes the majority of merge cases in 2007 (73.34%: 850/1159).

For example:

`L1024698|L0248601|S1235284|TRIM NAIL(S)`

Parenthetical plural form (S) is removed from NAIL(S) by luiNorm.2007 and thus "TRIM NAIL(S)" is merged into the same LUI (L0248601) of "Trim nail(s)".

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

The results of base form from Lexical tools base on the data of The SPECIALIST LEXICON. The base form of a word might be different from last year if there are new lexical records, modified records with new inflectional rules, or deleted lexical records associated with this word. Accordingly, this case is expected to be observed between releases and is considered as an enhancement. This case has 10.78% (125/1159) impact of merge cases in 2007.

For example:

`L0511581|L0511581|S1395160|Apis`

"API" is a new word (E0580402) in LEXICON 2007. Accordingly, "Apis" is uninflected to "api" and merged with string "API"

Also,

`B|C|L1918381|L0023726|S2234037|LINCOMYCINS`

Lexical record of "lincomycin" is modified by adding a new inflection rule 'variants=reg' in 2007. Accordingly, "lincomycins" is uninflected to "lincomycin" and merges with "lincomycin".

3). Canonical form:

There is an algorithm change in the generation of 2007 canonical forms. This change enhanced the luiNorm form by including all spelling variants and choosing the word with min. length from the same canonical class as the canonical form. This algorithm change introduced some of the merge cases. In addition, new lexical records (new EUI) and new spelling variants result in different canonical forms. 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
Algorithm change	151	82.06%	`cignolin, cygnolin, cygnoline`
New lexical records (EUI)	20	10.87%	`treehoppers`
New spelling variants	13	7.07%	`hypoalphalipoproteinemia`

Table 2. Percentage distribution of detail causes on Canonical forms

==========

- **Split:**
  - Summary:
    There are 25659 SUIs with 6671 LUIs splited to 13370 LUIs. Most of these splites are due to the enhancement on the Canonical form generation program. As mentioned above, luiNorm 2007 is enhanced by including all spelling variants and choosing the word with min. length from the same canonical class as the canonical form. In addition, lvg flow component of remove upper case parenthetical plural form (-f:rs) and new base forms from new data in Lexicon introduced less than 1% on the split cases. 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
Enhancement on the lvg flow –f:rs	18	0.27%	`diacetate(S)-isomer`
Base form (new data in Lexicon)	36	0.54%	`Stenoses, Aortic`
Canonical form (new algorithm & data)	6617	99.19%	`heartbeat`

Table 3. Percentage distribution of split causes

  - Split analysis:
    1). Lvg Flow component, –f:rs
    Most of this split happens when (S) appears in chemical terms. These 18 split cases might be a good study cases for enhancing the heuristic rules in –f:rs. However, due to the small impact of the split cases, this work is not recommended for 2008 release.

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

    3). Canonical form:
    The change of algorithm in the generation of 2007 canonical forms introduced almost 90% of the split cases. This software change is to enhance cases when the spelling of a word is the same as the spelling of an abbreviation.

    For example,
    `L0374476|L0374476|S0507619|R AW`
    `L0374476|L6119381|S0398663|RAW`

    “R AW” and “RAW” are split in 2007 while they belonged to same LUI in 2006. “R AW”, as a noun, has a spelling variant of “RAW” as acronym of “airway resistance”. Because of the same spelling, “raw” as an adjective, has the same LUI as “R AW” in 2006. This type of cases is taken care of in 2007. As for the cases of different canonical form caused by new EUI, we found some possible duplicated lexical record and will update our Lexicon database accordingly in 2008 release. The detail analysis on the causes of the split on canonical form and the results are shown in table 4.

Causes	Merge No.	Percentage	Example
Algorithm change	5936	89.71%	Raw
New lexical records (EUI)	193	2.92%	scopolaminebutylbromide
New spelling variants	488	7.37%	Anticoagulents

Table 4. Percentage distribution on detail causes on Canonical forms

==========

- **Split_Merge:**
  - Summary:
    There are 320 SUIs with 40 LUIs split_merge cases. These split_merges are due to the the enhancement on lvg flow component –f:rs, enhancement on the Canonical form generation program, and new base forms from new data in Lexicon. 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
Enhancement on the lvg flow –f:rs	23	57.50%	SKIN TAG(S)
Base form (new data in Lexicon)	8	20.00%	Stenoses
Canonical form (new algorithm & data)	9	22.50%	retrograde

Table 5. Percentage distribution of split_merge causes

Causes	Split_Merge No.	Percentage	Example
Algorithm change	7	77.78%	retrograde
New spelling variants	2	22.22%	interposition

Table 6. Percentage distribution of detail causes on Canonical forms

  - Split_Merge analysis:
    The cause of this case is the combination of above two (Split and merge). Potentially, terms in these 40 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 40 cases in 2007.

    1). Lvg Flow component, –f:rs

    Some of this split happens when (S) appears in chemical terms. These 23 split cases might be a good study cases for enhancing the heuristic rules in –f:rs. However, due to the small impact of the split_merge cases, this work is not recommended for 2008 release.

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

    Please refer to the discussion in merge cases.

3). Canonical form:

Please refer to the discussion in split and merge cases.