I. Summary

As one of our routine tasks for the annual Lexical Tools release, we analyze the results of LUI assignment based on OCCS reports of Merge, Split, and Split-Merge cases (from Ronny) to:

1) monitor the behavior of LuiNorm
2) enhance LuiNorm and its associated LVG flow components
3) fix Lexicon data for the next release (if any)

Based on the last year report, there is no software change suggested on the LuiNorm program. The only change for LuiNorm is the data from updated UMLS (atoms) and LEXICON. Accordingly, the results of LuiNorm should be the same between 2010 and 2011 releases except for records changed in LEXICON or UMLS. These changes are considered as system enhancement and are expected to happen every year with a relative small number resulting in merge, split and split-merge cases.

In our study, first, we observe the total number of merge cases (SUIs) decreased from 6,382 to 2,602 from 2010 to 2011 release. In addition, we found six of merge cases have same LuiNorm results. Theoretically, terms have same LuiNorm results should have same LUI assignment and no merge, split, or split-merge should be introduced between releases. These six cases were reported to OCCS and they were identified as new terms from OCCS investigation. Second, we observed the total numbers of all split cases (499) and split-merge cases (28) stay in the same level of small amount as expected. All split and split-merge cases have different LuiNorm results caused by different base forms or canonical forms. This small amount of changes is expected to happen every year and considered as system enhancement.

In conclusion, LuiNorm.2011 behaves very well and no software change requests or lexical records updates are found from this study. All three cases, merge, split, and split-merge, stay in a steady small amount of changes as expected. We also predict low magnitude of changes on merge, split, and split-merge for the next release (2012). We are pleased with the results and considering the algorithm of LuiNorm and OCCS LUI assignment programs are stable. Please refer to the web site for more details: http://lexlx1.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/designDoc/LifeCycle/test/luiAssign/index.html

II. Merge cases

- Summary:
  There are 2,602 SUIs with 1,053 Luis merged into 524 Luis. As expected, 98.85% (518/524) strings merged by different LuiNorm results. Table 1 shows the percentage distribution of all causes of merge cases. There are 87.78% and 11.07% of merge cases caused by the changes of canonical forms and base forms,

respectively. In addition, there are six merge cases that have same LuiNorm results, as shown in the examples field in Table 1. These merge cases are identified as new terms from OCCS investigation.

Causes of merge cases	Merge Lui	Percentage	Examples
Different canonical form (new data in LEXICON & UMLS)	460	87.78%	▪ bancroftus ▪ Phenofibrate
Different Base form (new data in Lexicon)	58	11.07%	▪ Cleanups ▪ IS ▪ Ribes ▪ Disgnostics
New terms, same LuiNorm result and merged	6	1.15%	▪ L0599233 ▪ L5583788 ▪ L5585353 ▪ L5587290 ▪ L5588656 ▪ L5591425

Table 1. Percentage distribution of merge causes

- Merge analysis:

  1). Different canonical form (changed data in UMLS - atoms or LEXICON):
  There is no software change in canonical forms generation program and LuiNorm in Lexical Tools. In other words, the results of canonical forms should be the same as the previous release except for changed terms in UMLS – atoms or LEXICON. These cases are expected to be observed every year and are considered as enhancements between releases. Below, we illustrate two examples in merge cases of these causes:

  Example 1 - New data in UMLS
  "bancroftus" is a new word in UMLS. "bancroftus" belongs to same canonical class of "bancrofti" and "bancroftu" in 2011 while it was not in 2010. Accordingly, "bancroftus" is merged with "bancrofti" into one LUI in 2011.

  Example 2 - New data in LEXICON
  "phenofibrate" is added as a new spelling variants of "fenofibrate" (E0301970) in the LEXICON 2011. Thus, "phenofibrate" belongs to same canonical class of "fenofibrate" in 2011 while it was not in 2010. Accordingly, "phenofibrate" is merged with "fenofibrate" into one LUI in 2011.

  2). Different base forms (changed data in Lexicon):
  The results of base forms from Lexical tools mainly depend on the data of LEXICON. The base form of a term might be different from last version if there

are new lexical records, modified or new inflectional rules, or deleted lexical records associated with this term. These cases are expected to be observed every year and are considered as an enhancement between releases. Two examples are illustrated as follows:

Example 1 - New data in a lexical record
A new inflectional rules, variants=reg, is added to the lexical record of “cleanup” (E0319808) in LEXICON 2011. According to this inflection rules, “cleanup” has inflectional variant of “cleanups” and merged with “cleanup” to the same LUI.

Example 2 - New lexical records
“I” is a new lexical record (E0701267) with part of speech of noun in LEXICON, 2011. According to the inflection rules (variants=metareg) in the lexical record, “I” has inflectional variant of “Is” and merged with “is”, “be”, “am” together to the same LUI.

III. Split cases

- Summary:
  There are 768 SUIs with 247 LUIs split to 499 LUIs. As expected, these numbers stay in the low magnitude. All these split cases are caused by different canonical forms (86.64%, 214/247) and base forms (13.36%, 33/247). As mentioned above in the Merge cases section, these cases are caused by the data change in UMLS or LEXICON. The total number of split cases is very small and we did not find any unexpected behavior of LuiNorm in the split cases. Table 2 shows the percentage distribution of each cause.

Causes	**Split No.**	**Percentage**	**Example**
Canonical form (new data in UMLS & Lexicon)	214	86.64%	▪ Monroe ▪ Meligramma guttatum
Base form (new data in Lexicon)	33	13.36%	▪ Chiropteras ▪ Omus ▪ hsides

Table 2. Percentage distribution of split causes

- Split analysis:

1). Different canonical form (changed data in UMLS - atoms or LEXICON):
As mentioned in the merge cases, different canonical forms can be caused by the data change in UMLS or LEXICON. Bellows, we illustrate an example of split cases caused by data change to result in different canonical forms.

Example – New lexical record in LEXICON
The term, “Monroe” was added to the LEXICON, 2011 (E0705321|noun) and did not exist in LEXICON, 2010. Accordingly, “Monroe” and “Monroes” are the only inflectional variants generated by this new lexical record and thus split from the canonical class of “monro” and results in a split case.

2). Different base forms (changed data in Lexicon):
Split cases can be caused from different based forms by the change of lexical records to result in different LuiNorm results. Bellows we illustrate an example of split cases caused by data change in LEXICON to result in different base form.

Example 1 - Deleted lexical records
“Chiropteras” is a lexical record (E0300873) with part of speech of noun in LEXICON 2010. It was deleted in LEXICON, 2011. The base form of “Chiropteras”, is changed from “Chiroptera” (generated by the facts, variants=reg, in 2010) to “Chiropteras” (generated by rules in 2011). Accordingly, different base forms result in a split case.

IV. Split-Merge cases

- Summary:
Split merge cases are the cases when some words split first, then merge (with others) again. There are 93 SUIs split into 28 LUIs and then merge into 18 LUIs in split-merge cases. The total number of split (28) merge (18) cases in 2011 is relatively small in the LUI assignment. Table 3 shows the percentage distribution of each cause. There are 39.28% (11/28), 14.29% (4/28), and 46.13% (13/28) caused by new canonical forms, base forms and no change but commit a split-merge case because of other split-merge cases, respectively.

Causes	**Split-Merge No.**	**Percentage**	**Example**
Canonical form (new data in UMLS & Lexicon)	11	39.28%	▪ HADS ▪ Daces
Base form (new data in Lexicon)	4	14.29%	▪ PDES ▪ GBS
Same luiNorm results, change because of other split-merge cases	13	46.43%	▪ PDE ▪ INTES

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 (2011), we did not found anything to enhance LuiNorm from these 18 cases. Please refer to the discussion in split or merge session for details.