

I. Summary

We are pleased with the results of Lui-Analysis:

- 1). the algorithm of LuiNorm is stable and no software change request is suggested
- 2). the OCCS Lui-Assignment program is stable
- 3). no term is identified for further investigation

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 Suresh) 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)
- 4) provide suggestions for OCCS Lui-Assignment program

Based on the last year report, there is no software change suggested on the LuiNorm program. The only change for LuiNorm is the data updates from the latest UMLS (atoms) and LEXICON. Accordingly, the results of LuiNorm should be the same between 2011 and 2012 releases except for the results from data changes in LEXICON or UMLS. These changes are considered as system enhancement and are expected to happen every year with a relative small number in merge, split, and split-merge cases.

In this study, first, we observed the total number of merge cases (SUIs) decreased from 2,602 to 1,361 from 2011 to 2012 release. In addition, we did not find any merge cases with same LuiNorm results in 2012 release while there were six such cases in 2011 release. Theoretically, terms have same LuiNorm results should have same Lui-Assignment and no merge, split, or split-merge cases are expected between releases. The expected result shows the high quality control in the 2012 release with no term needs further investigation. Second, we observed the total numbers of all split cases (117) and split-merge cases (15) stay in the same level of small amount as expected (actually it decreased compare to 2011 release). All split and split-merge cases have different LuiNorm results caused by different base forms or canonical forms. This small amount of change is expected to happen every year and considered as system enhancement.

In conclusion, LuiNorm.2012 behaves extreme well and no software change requests or lexical records updates are suggested from this study. All three cases, merge, split, and split-merge, stay in a steady small amount of changes as expected. We also predict similar low magnitude of changes on merge, split, and split-merge cases for the next release (2013). We are pleased with the results and considering the algorithm of LuiNorm and OCCS Lui-Assignment programs are stable. Please refer to the URL for more details: <http://lexlx1.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/designDoc/LifeCycle/test/uiAssign/index.html>

II. Merge cases

There are 1,361 SUIs with 674 LUIs merged into 336 LUIs. The total number of merge cases has drop about 50% from last year. All merge cases are caused by the changes of either canonical form (80%) or base form (20%). Table 1 shows the percentage distribution and examples of all causes of merge cases. No merge case with same LuiNorm results are found and thus no term needs further investigation. No further source analysis (from UMLS-atoms or LEXICON) is conducted because the expected results of small total number of merge cases.

Causes of merge cases	Merge LUIs	Percentage	Examples
Different canonical form (new data in LEXICON & UMLS)	269	80%	<ul style="list-style-type: none">▪ tremens▪ cruzy▪ zieve▪ kury
Different Base form (new data in LEXICON)	67	20%	<ul style="list-style-type: none">▪ bias▪ sachs▪ SAIDS▪ WSS
New terms with same LuiNorm result and merged for further investigation	0	0%	<ul style="list-style-type: none">▪ N/A

Table 1. Percentage distribution of merge causes

III. Split cases

There are 424 SUIs with 105 LUIs split to 117 LUIs. As expected, these numbers stay in a very low magnitude. All these split cases are caused by different canonical forms (62%) and base forms (38%). Please note that there is one term with the LUI of L1188409 changes in both Canonical form (“TPBE”) and base form (“TPBS”) and is counted in both causes of Canonical form change and base form change. Table 2 shows the percentage distribution and examples of each cause. As mentioned above in the Merge cases section, these split 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. No further source analysis (from UMLS-atoms or LEXICON) is conducted because the expected results of the small number of total split cases.

Causes	Split LUIs	Percentage	Example
Canonical form (new data in UMLS & LEXICON)	66	62%	<ul style="list-style-type: none">▪ Phenamine▪ Thioacetamid

			<ul style="list-style-type: none"> ▪ Demser
Base form (new data in LEXICON)	40	38%	<ul style="list-style-type: none"> ▪ Atlases ▪ Perinatologies ▪ Radiometries
Change in both Canonical form and Base form	1	NA	<ul style="list-style-type: none"> ▪ LUI: L1188409

Table 2. Percentage distribution of split causes

IV. Split-Merge cases

Split merge cases are the cases when some words split first, then merge (with others) again. There are 44 SUIs split into 15 LUIs and then merge into 10 LUIs in split-merge cases. The total number of split (15) merge (10) cases in 2012 is relatively small in the Lui-Assignment. Table 3 shows the percentage distribution and examples of each cause. There are 40%, 20%, and 40% caused by new canonical forms, base forms, and no change but commit a split-merge case because of other split-merge cases, respectively. 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 canonical algorithm to make canonical class covers more words. In this analysis, we did not found anything to enhance LuiNorm from these 15 cases.

Causes	Split-Merge LUIs	Percentage	Example
Canonical form (new data in UMLS & LEXICON)	6	40%	<ul style="list-style-type: none"> ▪ CNU ▪ CNI ▪ EOS ▪ EOE ▪ Hymenopterase ▪ BCHE
Base form (new data in LEXICON)	3	20%	<ul style="list-style-type: none"> ▪ Hymenopteras ▪ BCHS ▪ Teffs
Same luiNorm results, change because of other split-merge cases	6	40%	<ul style="list-style-type: none"> ▪ hymenoptera ▪ TEF ▪ EO ▪ Teff ▪ BCH ▪ CN1

Table 3. Percentage distribution of split-merge causes