Hi Brian and Suresh:

Excellent, excellent job! Thank you so much for sending us about these files and testing luiNorm 2005. As a matter of fact, we were kind of waiting for the feedback on the new feature of luiNorm2005. We expected more merge than split (less Canonical classes) due to the change on luiNorm2005 and Lexicon. As shown in your data, there are 7077 merge and 2219 split (as we expected). I went through some of the data and found the luiNorm behave as we expected. Bellows are the detail explanations if you are interested:

==========
1) Merge: 3087 Luis merge into 1536 Luis.

1-a). New feature of remove plural form of (s), (es), and (ies) in luiNorm.2005:
One of the major new feature of luiNorm.2005 is to remove plural forms in (s), (es), and (ies). This feature was requested by Stephanie Lipow and Guy Divita. As the first example shown in Merge file, "Cramp(s), abdominal", is normalized to "abdominal cramp". Other similar examples are:

"Amputation of limb(s)"
"Tuberculosis;bone(s)"
"Polyp(s);cervical"
"Transplantation of testis(es) to thigh"

There are:
- 1055 cases of removing plural form of (s)
- 2 cases of removing plural form of (es)
- none for (ies)

Total, 1057, takes about 68% (out of 1551) of merge case.

1-b). Combined spelling variants records in Lexicon.2005:
One of major change in Lexicon.2005 is to combine records which are spelling variants. For examples, "abscess" and "abcess" are two separate records before and it is combined as one record with specifying spelling variants relationship in 2005 Lexicon. Most of the merge data from the rest 32% are the results of this improvement of Lexicon.

1-c). Improved algorithm:
Improved algorithm for normalization and canonical classes generation in 2005 (for cases of s followed by punctuation). For example, "Self-Monitorings, Blood Sugar, Blood" is normalized to "blood monitor self sugar", while in previous version, it was normalized to "blood monitorings self sugar". Some merge data are caused by this improved canonical algorithm.

1-d). Updates on lexical records in Lexicon.2005:
For example, Lexical record of "Progeria" is added new attribute of "variants=reg" in Lexicon 2005. Accordingly, "Progerias" is now

normalized to "Progeria" while it had the normalized form as "Progerias" (itself) in the previous year. "Hirudin", " Imidazoline", etc. are some another examples for this case.

==========
==
2) Split: 940 Luis split to 1884 Luis.

2-a). remove plural form of (s), (es), and (ies) in luiNorm 2005:

This new feature only remove the pattern with lower case of s inside the parenthesis, not the upper cases. The main reason for this is the concern of not to remove (S) in Chemical terms by accident. There are 750 luis (out of 944, 80%) split because of this reason. I will need more study and data test before I feel comfortable to enhance this feature to take care of upper case.

An example for split in above case is:

- ADJUSTMENT OR REVISION OF EXTERNAL FIXATION SYSTEM REQUIRING ANESTHESIA (EG, NEW PIN(S) OR WIRE(S) AND/OR NEW RING(S) OR BAR(S))

is split from

Adjustment or revision of external fixation system requiring anesthesia (eg, new pin(s) or wire(s) and/or new ring(s) or bar(s))

Due to the remove (s), not the (S).

2-b). More lexical records in Lexicon:
There is a heuristic rule in lvg uninflection flow component while using rule (not the fact). Which is to drop uninflected forms derived by rules if they are in the Lexicon. For example, "Aujeszkys" is uninflected to "Aujeszky" by rule. However, "Aujeszky" is in Lexicon 2005. Thus, it is dropped. However, "Aujeszky" is not in the previous year Lexicon. Accordingly, "Aujeszkys" is uninflected (normalized) into "Aujeszky" in 2004.

This makes some Luis split. Such as:
- "Aujeszkys Disease Virus" split from "Aujeszky Disease Virus"
- "Intraoral Electrogalvanisms" split from "Intraoral Electrogalvanism"
- "Gravis, Myasthenia" split from "Gravi, Myasthenia"
...

Also, "electrolyses" is now uninflected and canonical into "electrolyse" class because of a new record "electrolyze" (E0235411) is added in 2005 Lexicon.

Similarly, "deermice" is uninflected into "deermouse" in 2005 while it was uninfelcted into "deer mouse" in the previous year.

These cases take most of the split from the rest 20% of split.

2-c). Combined spelling variants records in Lexicon.2005:
As stated in 1-b). records are combined if they are spelling variants in Lexicon 2005. For example, "autolyse" is in the same records as "autolyze" and "autolise" in 2005 Lexicon. This results in normalizing "Autolyses" to "autolise" in 2005 while "Autolyses" is normalized to "autolysis" in 2004. This is the "is/es" problem as Suresh observed. Thank you, Suresh!

Also, due to the new format of spelling variant and irreg in Lexical record. It precisely points out the one to one relationship in the irreg case. For example, "clubfeet" is uninflected into "clubfoot" in 2005 while it was uninfelcted into "club foot", "club-foot", and "clubfoot" in the 2004. This results in different canonical classes.

2-d). Updates on lexical records in Lexicon.2005:

For example, "methylhistidine" (E0401123), it's attribute of "variants=reg" has been removed in 2005. In other words, "methylhistidines" is not the plural form of "methylhistidine" in Lexicon 2005 (while it was in 2004). Thus, "methylhistidines" is uninflected to "methylhistidines" by rule in 2005 while it was uninflected to "methylhistidine" by fact in 2004.

This also results in the other problem Suresh observed. "lower" and "lour" are spelling variants (E0038088, verb) in 2005 and thus they are in the same canonical class along with "lowered". However, they are not spelling variants as in verb in 2004. Thus, lowered was canonical to lower in 2004.

==========
=
3) Split-Merge:

I didn't spend too much time of this file since it is not too many cases (336). However, there are 169 (50%) of them are caused by remove (s), (es) and not remove (S), (ES).

I would expect most of the rest was caused by:
- more records of Lexicon
- combined records by spelling variants in Lexicon
- new format of irreg on spelling variants in Lexicon
- data update of Lexicon

Last but not the least, I would like to thank Brian and Suresh for helping us to test the new features of luiNorm.2005. Also, I hope my explanation make sense. Thanks!

-- Chris

==========

Chris and Suresh,

Each time LVG versions change we run a sample LUI re-assignment to test the new algorithm. The result of this is three reports showing the splits, merges, and split-merge cases resulting from the new LVG. Each report has the form OLD_LUI|NEW_LUI|SUI|STRING and it shows how LUIs from the previous version were changed in the current version.

The three reports are attached. There were 1536 merged LUIs, 940 Split LUIs, and a couple hundred of the complex split-merge case.

Following are 1 example from each report:

split_merge.dat

Here L0016358 split partially and merged with L1222595 which also split partially:

```
L0016358|L0016358|S0031164|Dental Fluorosis|
L0016358|L0016358|S0042137|Fluorosis, Dental|
L0016358|L0016358|S0220882|Fluorosis dental|
L0016358|L0016358|S0371828|FLUOROSIS DENTAL|
L0016358|L0016358|S0481037|Dental fluorosis|
L0016358|L0016358|S4032308|Dental fluorosis <1>|
L0016358|L0016358|S4032309|Dental fluorosis <2>|
L0016358|L0016358|S4071834|dental; fluorosis|
L0016358|L0016358|S4082970|fluorosis; dental|
L1222595|L0016358|S0031163|Dental Fluoroses|
L1222595|L0016358|S0042136|Fluoroses, Dental|
L1222595|L1222595|S0031162|Dental Fluorose|
L1222595|L1222595|S0042135|Fluorose, Dental
```

merge.dat

Here L0000729 and L1697876 merged into L000072.

```
L0000729|L0000729|S0009058|Abdominal Cramps|
L0000729|L0000729|S0219866|Cramp, abdominal|
L0000729|L0000729|S0219867|Cramps, abdominal|
L0000729|L0000729|S0226111|ABDOMINAL CRAMP|
L0000729|L0000729|S0351956|ABDOMINAL CRAMPS|
L0000729|L0000729|S0353641|Abdominal Cramp|
L0000729|L0000729|S0353650|Abdominal cramps|
L0000729|L0000729|S0361688|CRAMP ABDOMINAL|
L0000729|L0000729|S0364797|Cramp, Abdominal|
L0000729|L0000729|S0364801|Cramps, Abdominal|
L0000729|L0000729|S1616770|Abdominal cramp|
L0000729|L0000729|S1616771|Cramp abdominal|
L0000729|L0000729|S1911033|Cramps;abdominal|
L1697876|L0000729|S1911042|Cramp(s);abdominal|
L1697876|L0000729|S1932181|Cramp(s), abdominal|
```

split.dat

Here L0004315 split into L0004315 and L5310923
L0004315|L0004315|S0016881|Aujeszky Disease Virus|
L0004315|L0004315|S0016884|Aujeszky's Disease Virus|
L0004315|L0004315|S0033563|Disease Virus, Aujeszky|
L0004315|L0004315|S0098604|Virus, Aujeszky Disease|
L0004315|L0004315|S0098605|Virus, Aujeszky's Disease|
L0004315|L0004315|S0594335|Aujeszky's disease virus|
L0004315|L5310923|S0016887|Aujeszkys Disease Virus|

Brian