Subject: Changing the arrangement of base forms in the lexicon.

The citation form of a lexical record is the form that fills the "base=" slot in a lexical unit record. We have always stressed that this form was arbitrarily chosen. Often it has been the first form we saw when the entry was created. Other morphological base forms have fallen into the spelling variants slots in no particular order.

We have decided that it would be beneficial to make the choice of citation form predictable if not less arbitrary and to predictably order the spelling variants as well. In future versions of the lexicon the citation form will be chosen by algorithm and spelling variants will be sorted in a predictable order.

The base forms will be ordered so that the first (citation from) will be pure ASCII; forms without punctuation characters will precede those with punctuation characters; shorter forms will precede longer forms and alphabetic order will be used to reduce any remaining disorder.

This Web page describes the sorting order for base forms (citation form and spelling variants) in detail with examples:

http://lexlx1.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/lexRecord/content/baseOrder/index.html

Rationalizing the order of spelling variants and the selection of the citation from will allow us to automatically check and correct cross references, e.g. those between Acronyms and their expansions or verbs and their nominalizations.

Chris has already implemented the change and reports that only 6 LVG flows are affected. The change affects only 7.15% of lexical records and no unexpected issues were found in his testing. Since we're replacing one meaningless ordering of base forms with another meaningless ordering we think the changes should not affect the workings of the affected LVG flows.