Lexicoder Sentiment Dictionary
Codebook

Lori Young
University of Pennsylvania
lyoung@asc.upenn.edu

Stuart Soroka
University of Michigan
ssoroka@umich.edu

August 2015

This codebook is part of a zip file distributed from lexicoder.com — a file which includes the Lexicoder Sentiment Dictionary (LSD) alongside related files and information. In addition to the current file, the package includes:

LSD2015.lc3	The Lexicoder Sentiment Dictionary, August 2015 release, formatted for use with Lexicoder 3.0 (but easily reformatted for other software as well).
LSD2015_NEG	The negated version of the Lexicoder Sentiment Dictionary, formatted for use with Lexicoder 3.0 (but easily reformatted for other software as well).
LSDpreprocess2015	A plain text file (.txt) of the preprocessing modules used alongside the LSD and a folder containing script files (.scpt) to implement preprocessing.

Each of these files is discussed further below. The .lc3 files are readable in any plain text software; they are is designed for use with the Lexicoder software, which is freely available for research purposes at http://www.lexicoder.com. Note that the August 2015 release of the LSD is no different from the 2011 release, except that dictionaries are formatted for the newest version of Lexicoder. Updates to the dictionaries and preprocessing modules are forthcoming, however.

The objectives, development and reliability of the dictionary are discussed in detail in Lori Young and Stuart Soroka, 2012, "Affective News: The Automated Coding of Sentiment in Political Texts," *Political Communication* 29: 205-231. Please cite this text when using the Lexicoder Sentiment Dictionary and related resources.

The LSD

LSD2015

The LSD dictionary contains two large valence categories for positive and negative sentiment, comprising roughly 4,500 unique entries. The .lc3 file is ready for use with Lexicoder 3.0. For information on use of the Lexicoder software, please download the User's Manual from the Lexicoder website.

LSD2015_NEG

Many users will find the standard version of the LSD adequate for text analysis. Nonetheless, Young and Soroka (2012) did find a small, but non-negligible increase in performance when accounting for negations. Thus, we have included LSD2015_NEG.lc3. The format of this file is identical to the LSD2015, as described above, however the entries are negated (e.g. "happy" is "not happy"). To maximize the utility of the negated version of the LSD we strongly recommend running the negation preprocessing module first, as it standardizes a number of negation phrases. For a full description of the development of the negated version of the dictionary and the related negation preprocessing module, see Young and Soroka (2012).

SCORING

The output from the Lexicoder Software is described in detail in the User's Manual. The LSD is typically scored using Lexicoder's Dictionary Counter, which produces a frequency count of dictionary terms in each category for every unit of analysis. Lexicoder also includes a Word Counter; LSD scores can be calculated as a percentage of words in the full text. (There are other options too, of course, though we will use the difference between positive and negative words, as a proportion of all words in the text, as an example below.)

Accounting for negations in the text requires running both the LSD2015 and LSD2015_NEG dictionaries. Negated terms will be counted by both dictionaries, e.g. "not happy" will be captured as "happy" in the "POSITIVE" list in LSD2015, and will be captured as "not happy" in the NEG_POSITIVE (negated positive) list in LSD2015_NEG. Generating a final score requires a quick calculation: a negated positive word in LSD2015_NEG should be subtracted from the positive word score from the LSD2015 dictionary, and added to the negative word score.

Consider the following example: a text produces a count of 10 positive words and 5 negative words using the LSD2015 dictionary. The text is 100 words long, so one "net tone" measure could be (10-5)/100=.05. Running the LSD2015_NEG dictionary, however, we find that there are 3 negated positive words; or, rather, that 3 of the positive words may actually be negative. We need to subtract those from the initial positive count, and add them to the negative count. So the new "net tone" measure is (7-8)/100=-.01. More generally, we can adjust the initial counts using the negation dictionary as follows:

Number of positive words:
LSD2015[Positive] – LSD2015_Neg[Neg_Positive] + LSD2015_NEG[Neg_Negative]

Number of negative words:
LSD2015[Negative] - LSD1015_Neg[Neg_Negative] + LSD2015_NEG[Neg_Positive]

Note that the LSD is best suited for analyzing large bodies of text. All tests were conducted on large samples at the article-level and reliability improves with the number of words analyzed. We do not recommend analyzing text at the sentence level, as it is much less likely to be reliable.

Sentence-level characteristics of the text can however be taken into account using preprocessing of text described in the Lexicoder User's Manual. (This can be done in other software as well of course.) It is also possible to the Hierarchical Dictionary Count (hdc) function in Lexicoder 3.0 to calculate the occurrence of words in one dictionary category appearing in the same sentence as words in another dictionary category. For example, the number of positive or negative terms in the same sentence as mentions of a given political actor has be used to measure the tone of media coverage toward a given party or leader. Using the hdc function, although the module analyzes sentence-level features of the text, the output is aggregated at the article-level. A full description of use of this module is available in the Lexicoder User's Manual.

Preprocessing

The reliability of the LSD is significantly improved if text is preprocessed before it is coded. The general purpose of preprocessing is to standardize the text to improve the scope of coverage and to remove non-sentiment homonyms from the analysis. The preprocessing of over 1500 words and phrases facilitates basic word sense disambiguation

and the contextualization of many commonly used sentiment words and phrases. A full description of the development of the preprocessor, including a number of examples, is provided in Young and Soroka (2012).

The preprocessor consists of the following four process modules:

1-punc	The punctuation module standardizes and/or removes punctuation
2-caps	The caps module removes capitalized words (i.e. proper nouns, which by definition have no tone)
3-negation	The negation module standardizes a set of commonly used negation phrases (Note that this module is particularly critical if you are using the negation dictionary, which relies on these contracted negations)
4-dict	The dictionary module removes false hits for dictionary entries, that is, topical, multi-toned or non-tonal instances of dictionary terms

The preprocessing used in Young and Soroka (2012) was written and implemented in Applescript and run on a Mac using TextWrangler, a free plain-text editor with a powerful find-and-replace mechanism. TextWrangler is an easy-to-use program, available to download at http://www.barebones.com/products/textwrangler/.

The preprocessor is distributed here as a set of Applescript (.scpt) files in the folder LSDpreprocess2015, and ready for use in TextWrangler. The files are labeled and ordered for each process module. Note that many entries in the preprocessor are contingent on those before and should thus be applied in order.

A tab-delimited .txt file of the entire preprocessing routine is also included in the package for reference purposes. Entries that require more than basic search and replace commands are in square brackets. The file contains four columns: a sort ID, the module type, the word or phrase to be replaced, and the word or phrase to replace with.

Comments and queries are very welcome, at lexicoder@me.com.

Enjoy!