Dev IL Pack Components and Structure

Directory Structure

Note: "LANG" is a variable to be replaced by the 3-letter iso code for the incident language

README.txt
data/
  monolingual_text/
  translation/
    found/
      eng/
        ltf/
        psm/
      sentence_alignment/
      LANG/
        ltf/
        psm/
    comparable/
      eng/
        ltf/
        psm/
      clusters/
      LANG/
        ltf/
        psm/
docs/
  categoryII/
  categoryI_dictionary/
dtds/
tools/
  encoding/
  ltf2txt/
  twitter-processing/

Data Volumes

Monolingual text – 225Kw (45% NW, 33% DF/WL, 22% SN)
Parallel text – 300Kw (33% each NW, DF/WL, SN; can substitute 300Kw comparable for 100Kw parallel)
Parallel dictionary – 10K lemmas
Other resources ("Category II") – 5 of 8 resource types: parallel IL - > non-English dictionary, monolingual IL dictionary, monolingual IL grammar book, parallel IL - > English grammar book, monolingual IL primer, monolingual IL gazetteer, parallel IL - > English gazetteer, English gazetteer for incident region

Note: Minimum target to be exceeded by 500% target for **one** of monolingual text, parallel text, or parallel dictionary

Documentation

The following items should be included in all dev IL docs/ directories:

- categoryI_dictionary/ – directory containing dictionary (or pointer to dictionary)
- categoryII/ – directory containing (pointers to) all category II resources
- source_codes.txt – 4 columns: genre, source code, source name, (base) url
- twitter_info.tab – 4 columns: file name, "network_id" (Tweet ID), md5sum, author_id
- urls.tab – list of all source docs with document urls