README

1. Publication Title: ACE 2005 Multilingual Training Corpus (English SpatialML annotation)
2. Authors
    a. Christy Doran (point of contact, cdoran@mitre.org)
    b. Inderjeet Mani (imani@mitre.org)
    c. Seamus Clancy (sclancy@mitre.org)
    d. Janet Hitzeman
3. Data type: text
4. Data sources: The source documents come from the Linguistic Data Consortium (catalog ID: LDC2006T06); the documents happen to be those used by the Automatic Content Extraction program (ACE'2005). The catalog entry is http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T06. This corpus provides annotation of these files with SpatialML (http://sourceforge.net/projects/spatialml).
5. Project: Spatio-Temporal Information Extraction and Reasoning
6. Applications: natural language processing, spatial reasoning
7. Languages: English
8. Special license: See Distribution Agreement in the **doc** directory
9. Grant number and funding agency: 05MSR119, MITRE-sponsored research
10. Copyright: LDC and The MITRE Corporation
11. Description of the corpus structure and data attributes
    a. Data type: Text, in both in-line xml format and aif format
    b. 428 sgml files, 428 corresponding aif files, 428 corresponding gaz-deref, and 428 sgm.dtdvalidated files. The gaz-deref files contain multiple gazetteer references when they exist for a single location; these different gazrefs sometimes correspond to slightly different latlongs. The gazateer references in the .sgm files are only to IGDB. The sgm.dtdvalidated files do not contain document structure tags (such as <DOC>, <TXT>) that would prevent them from being validated with the SpatialML DTD. These files total 22624650 bytes uncompressed.
    c. The *.sgm and *.gaz-deref files are in xml format and the .aif.xml files are in aif format that can be manipulated the Callisto annotation tool using the SpatialML plugin version 0.10.2.
    d. Counts of unique words can be found in the file doc/ldc_wordcount.csv and include all words that aren't part of XML markup (so no tag names, no attribute names or values, etc). Unique words are counted by comparing case insensitive transformations with preceding and trailing punctuation stripped off. "Words" consisting solely of punctuation are discarded. In all the corpus contain 210065 total words and 17821 unique words.
    e. Directory contents
        i. The **doc** directory contains this README file plus a Word file containing annotation guidelines, the Distribution Agreement in pdf format and a document in csv format listing the unique word counts for the individual files plus their total.

ii. The **data** directory contains the files described in section 11.b

iii. The **dtd** directory contains the SpatialML DTD.

12. Quality control: This data is validated via the SpatialML DTD which can be found in the **dtd** directory.