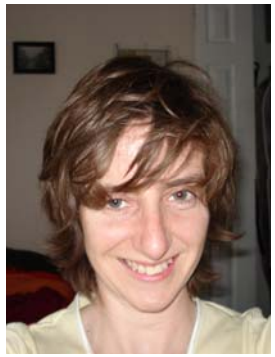
LIRICS Data Categories for Semantic Annotation



Harry Bunt Amanda Schiffrin
Tilburg University

Paris, 10 May, 2007

Background for non-ISO/SIGSEM experts

- 2002: Establishment of ISO TC 37/SC 4: Terminology and Language Resources
- 2003: Establishment of ACL-SIGSEM Working Group on the *Representation of Multimodal Semantic Information*
- 2004: Estalishment of ISO TC 37/SC 4/ TDG 3: Semantic content
- 2005: Start of European project LIRICS: Linguistic Infrastructure for Inter-operable Resources and Systems

ISO TC37/SC 4/TDG 3

- Aim: build data categories (= *well-documented, generally agreed descriptions of concepts in online registry*) for semantic annotation and representation, in particular for:
  - temporal information
  - semantic roles
  - dialogue acts
  - discourse relations
  - referential relations

ACL-SIGSEM

- ACL-SIGSEM Working Group for the Representation of Multimodal Semantic Information.

Aim:

To provide a platform for researchers in computational semantics and multimodal dialogue to discuss, evaluate and support possible future work on multimodal semantic content.

LIRICS

- Aim: build data categories for ISO certification, for syntactic, morphosyntactic, and semantic annotation, and for lexicon construction.
- WP4, Semantics: selection of TDG 3 work items:
  - referential relations
  - dialogue acts
  - semantic roles
  - temporal information - *new ISO project SemAF, Part I, Time and Events, based mainly on TimeML (James Pustejovsky, Bran Boguraev, Harry Bunt, Kiyong Lee, Nancy Ide) started in 2006*

ACL-SIGSEM Activities

Activities of the Working Group for the Representation of Multimodal Semantic Information:

- Inaugural meeting, Tilburg, the Netherlands January 2003, at IWCS-5
- Discussion meetings:
  - Nancy, September 2003
  - Lisbon, May 2004, at LREC'04
  - Tilburg, January 2005, at IWCS-6 (with ISO)
  - Marina del Rey, Cal., April 2006 (with ISO)
  - Tilburg, January 2007, at IWCS-7 (with ISO)

TDG 3 Activities

- Inaugural meeting: Lisbon, May 2004, at LREC'04 (joint meeting with ACL-SIGSEM WG)
- Project meetings:
  - Tilburg, January 2005 (with SIGSEM)
  - Paris, May 2005
  - Warsaw, August 2005
  - Jeju, Korea, January 2006
  - Marina del Rey, April 2006 (with SIGSEM)
  - Boston, October 2006, on SemAF/Time
  - Tilburg, January 2007 (with SIGSEM)

Most Recent TDG 3 Activities

- *Temporal information:* new ISO project: SemAF, Part I, Time and Events, based mainly on TimeML (James Pustejovsky, Bran Boguraev, Harry Bunt, Kiyong Lee, Nancy Ide)
- *Dialogue acts* and *Semantic roles:* NWIP to ISO, may be formulated in the course of 2007 (by Harry Bunt, David Traum ??) the latter based mostly on FrameNet and PropBank.

ISO/LIRICS Work

- Development of methodology for designing semantic annotation schemes
- Design of metamodels for focal areas of semantic annotation
- Design of data categories for core semantic roles, dialogue acts, temporal concepts and coreference relations, documented in LIRICS deliverable D4.2
- Publication of accompanying studies at LREC, ACL-SIGDIAL, IWCS, DECALOG, ...

Semantic Annotation Schemes 1: Time

Temporal annotation:

- Data categories in LIRICS deliverable D4.2 inspired primarily by TimeML and OWL-Time proposals
- “Temporary” data categories, hopefully useful in SemAf/Time project.

Semantic Annotation Schemes 2: Reference

Reference annotation:

- Data categories based primarily on studies by Laurent Romary and Susanne Alt, developing work by Cruse (1986) and van Deemter & Kibble (2000)
- Besides specific datcats for reference, also quite general datcats, such as /cardinality/, /countability/, /definiteness/...

Semantic Annotation Schemes 3: Dialogue acts

Dialogue Act annotation:

- Data categories based on comprehensive comparative studies of approaches to dialogue interpretation
- Taxonomy of 62 data categories in LIRICS deliverable D4.2 based primarily on DIT, DAMSL and Allwood's work

Semantic Annotation Schemes 4: Semantic roles

Semantic roles

- Data categories based on comparative study of FrameNet, PropBank, EAGLES, and literature: Dowty, Jackendoff...
- Current choice of 29 semantic roles determined primarily by:
  - generality of concepts
  - “definability”

LIRICS Current Work: Test Suite Construction Task

- Test suite: small manually annotated corpus, illustrating (and validating) the set of semantic data categories.
- For English, Dutch, Italian, Spanish, French and German: test suites for
  - Semantic roles
  - Dialogue acts
  - Referential relations

  Not for temporal information (in view of TimeBank)

Organization of LIRICS Test Suite Construction Task

	per language	*in total*
Semantic roles	500 sentences	3000 sent.
Dialogue acts	500 utterances	3000 utt.
Reference	500 sentences	3000 sent.

Organization of LIRICS Test Suite Construction Task (2)

- Dialogue acts:
  - 5 information-seeking dialogues of 50 utterances = 250 utterances
  - 5 assistance-seeking dialogues of 50 utterances = 250 utterances
- Reference:
  - 6 texts of 25 sentences = 250 sentences
  - 4 dialogues of 50 utterances = 250 utterances
- Semantic roles:
  - 125 PropBank/Penn Treebank (WSJ) sentences
  - 125 FrameNet sentences
  - 4 reference-annotated texts = 150 sentences
  - 2 reference-annotated dialogues = 100 sentences

Organization of LIRICS Test Suite Construction Task (3)

- For semantic role annotation: use of GATE software, with specification files for annotation schemes; separate annotation guidelines; output in XML
- For reference annotation: use of PALINKA software if annotation scheme easily modifiable, else use GATE; output in XML
- For dialogue act annotation: use of ANVIL software, with annotation scheme and annotation guidelines added; output in XML

Latest Developments

English (Tilburg): well under way for semantic roles and dialogue acts; getting ready for starting on reference; additional contribution from DFKI

Dutch (Tilburg: well under way for dialogue acts; started for semantic roles; reference to follow

Italian (Pisa): started for semantic roles and dialogue acts; reference to follow soon

Spanish (Barcelona): started for semantic roles; intention to contribute for dialogue acts and reference

French: unclear.

Interannotator Agreement

- Annotation of texts by at least three different annotators
- First phase: collaborative
- Second phase: individual
- Annotations from second phase will allow comparison of consistency of annotation, and the applicability of semantic data categories