Modeling Lexical Entries in Bilingual Dictionaries

—Or—

Exegeting the UML Model

Mike Maxwell

Linguistic Data Consortium

Three Levels of Abstraction

- File formats
- Data models
- Ontologies

Conceptual Structure vs. Views

- Data model = Conceptual/ Underlying structure
- View = layout, formatting
- Examples of views:
  - Page layout
  - Definition numbers
  - Alphabetization
  - Filtered subsets

Conceptual Structure vs. Views

- Spanish-English and English-Spanish sides of bilingual dictionary: View
- Spanish lexical entries, English lexical entries, and relations between them: Underlying structure

UML Models

- What is UML?

  "The Unified Modeling Language™ (UML) is the industry-standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems. It simplifies the complex process of software design, making a 'blueprint' for construction."
  (http://www.rational.com/uml/index.jsp)
- Blueprint language
- We'll use small subset

UML Models

- Objects
- Classes
- Attributes
- Links
  - Composition
  - Association
- Class hierarchy

UML Models

- Normalization
  - Data item appears *once*
  - Attribute (‘field’) holds *one* type of data
- Strings
  - MultiUnicode
  - MultiString

SIL-developed Model

- Bilingual lexicon
  (one-way: full information for vernacular language only)
- Developed for LinguaLinks
- Modified for Fieldworks
- Embedded in larger model of language description (http://fieldworks.sil.org/ModelDoc/)

Lexicon

- Front matter, appendices, …
- Lexical entries
  - Lexemes (stems, roots, words)
  - Affixes
  - Larger constructs (idioms etc.)

Lexical Entry

- Kinds of lexical entries
  - Major Entry
  - Subentry
  - Minor Entry

Major Entries

- LexMajorEntry
- For morphemes and non-compositional word-level “things”
  - Stems, roots, affixes
    (not a theoretical statement!)
  - But citation forms can be words

Subentries

- LexSubentry
  - Subclass of LexMajorEntry
- For multi-morphemic constructs:
  - Derivatives
  - Compounds
  - Idioms
  - Sayings
  - Phrasal verbs

Subentries (cont'd)

- Points to morphemes (etc.) of which it is composed
- Does not "belong" to morphemes (LexMajorEntries) of which it is composed

Minor Entries

- LexMinorEntry
  - Subclass of LexMajorEntry
    (but usually much simpler)
- For irregular forms (*oxen*, *been*, *went*)
- Belong to a LexMajorEntry
  (but alphabetization is a view!)

Parts of Lexical Entries

- Lexica est omnis divisa in partes tres (plus a label):
  - Citation form (= the label)
  - Forms
  - Morphosyntactic information
  - Senses
- No provision for etymology

Parts of lexical entries: Citation Form

- = Lemma, Headword, Canonical Form
- CitationForm attribute
- multiUnicode

Parts of lexical entries: Forms

- Pronunciations
  LexPronunciation (written form + sound)
- Allomorphs
  MoForm (written form, morph type, phonological context…)
- Underlying Form
  MoForm

Parts of lexical entries: Morphosyntactic Information

- MoStemMsi (for Stems/ Roots, whether bound or free)
  - Part of speech
  - Inherent morphosyntactic features
  - Inflection class (= paradigm/ declension)
  - Exception features

Parts of lexical entries: Morphosyntactic Information

- MoInflectionalAffixMsi (for Inflectional Affixes)
  - Morphosyntactic features
  - Exception features

Parts of lexical entries: Morphosyntactic Information

- MoDerivationalAffixMsi (for Derivational Affixes)
  - From/ to POS
  - From/ to morphosyntactic features
  - From/ to inflection classes
  - From/ to exception features

Parts of lexical entries: Senses

- LexSense:
  - Definition
  - Gloss
  - Scientific name
  - Pictures
  - Example sentences
  - Sub-senses (more LexSense objects)

Parts of lexical entries: Senses

- LexSense (cont'd):
  - Morphosyntactic information: *points to* a 'MorphosyntaxInfo' object
  - This MorphosyntaxInfo' object can be shared among different senses of the same LexEntry:
    run = to jog
    run = to go (to the store)
    (both can be nouns or intransitive verbs)

Parts of lexical entries: Senses

- LexSense (cont'd):
  - Use of shared 'MorphosyntaxInfo' object allows flexibility via views:
    The particular way in which definitions and other features of the dictionary article are presented comprise the macrostructure. Are definitions arranged by part-of-speech?… (Landau, *Dictionaries: The Art and Craft of Lexicography*, p. 99)
  - A view!

Parts of lexical entries: Senses

- LexSense (cont'd):
  - *Points to* set of 'ReversalIndexEntry' objects
    - Can be shared among senses belonging to the same or other LexEntries
    - Many-to-many relation between LexSenses and ReversalIndexEntries

Parts of lexical entries: Senses

- ReversalIndexEntry:
  Impoverished LexEntry
  - Name (= citation form)
  - POS
  - Sub-entries
    Allows for reversal entries like:
    Green (adj.)
    to be green: yax

Relationships among Senses: Synonyms

- LexSimpleSet
  One set per group of synonyms
  (asymmetry in model?)
- 'Members' = LexSetItems, in turn pointing to a LexSense
  (LexSetItems are a throw-away class?)

Relationships among Senses:
Antonyms and other Binary Relations

- LexPairRelations, owning sets of LexPairs
- Allows:
  - Directed relations (e.g. individual-group)

  or
  - Undirected relations (e.g. antonyms)

Relationships among Senses: Part-Whole, Generic-Specific

- LexTreeRelations, owning sequence of LexTreeItems
- Outline structure:
  (animal (mammal (dog cat))
  (reptile (snake turtle)))

Relationships among Senses: Scales

- LexScale
  (relation not specified: asymmetry in model)
  - Negative-neutral-positive scales
    (*tiny, small; medium; big, huge*)
  - Positive (or neutral) scales
    (*inch, foot, yard, furlong*)
    (*January, …December*)

Dialects

- Q: What can vary between dialects?
- A: Anything

Dialects

- Modeling dialects
  - Separate encodings
  - Separate lexicons
  - Mark objects for dialect (what level of granularity?)