Principles of the GOLD Ontology & Conversion of GOLD to DCIF

Presenters: Anthony Aristar, Evelyn Richter

Outline

- The Structure of the GOLD Ontology
  - History of GOLD
  - Why an Ontology Was Chosen
  - Structure of GOLD
  - One Advantage of an Ontology
- GOLD to DCIF
  - Refining GOLD
  - Mapping the Content of GOLD to DCIF
  - Mapping the Structure of GOLD to DCIF
  - Summary & Outlook

How GOLD began

- GOLD had its origin in the very first EMELD workshop, held at Santa Barbara in 2000.
- At this stage there was no consensus on how to deal with linguistic markup.
- Most people wanted linguists to use a single set of markup categories, and try and “force” everyone to use it.

The Solution

- Since the idea that making academics do anything they didn’t want to was equivalent to trying to herd a group of Siamese cats, the workgroup on markup (which included Gary Simons & Terry Langendoen) understandably came to the conclusion that no one would use the same markup
- So, let everyone use the markup they wanted, and build a set of concepts that everyone could link to automatically

Background

- Behind this was the fact that both Gary and Terry had been part of the group that planned the TEI.
- Using the ideas behind the TEI, you can use automatic methods to interpret tags.
- There was a general feeling that something inspired by the TEI would be used, though TEI had been intended for format, not content.

Aug 4, 2010 RELISH Workshop 5

The Beginning of GOLD

- Just then, the Suggested Upper Merged Ontology (SUMO) emerged.
- This was remarkably innovative for its time… the idea was based on the idea that with the right structure — an ontology — machine reasoning was possible.
- It was thus proposed that an ontology built on SUMO serve as “interlanguage” for translation among markups.

Aug 4, 2010 RELISH Workshop 6

Such an Ontology would:

- Have a concept hierarchy
- Have an enriched taxonomy
- Act as an interlingua

And finally it would:

- Eliminate the need for a gold standard of linguistic terminology

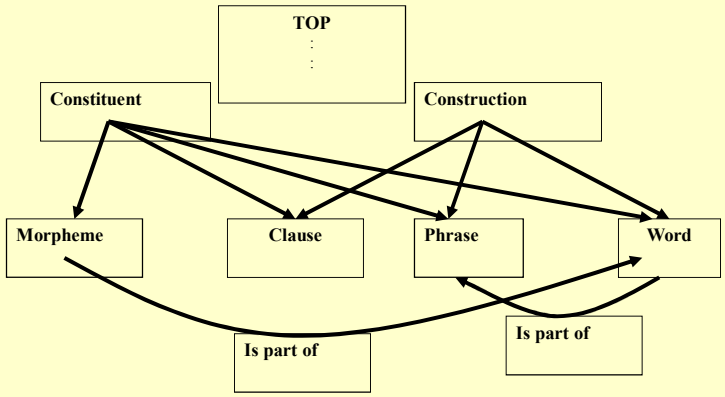
The Structure of GOLD

- It follows the standard ontological structure you would expect:
  - It is a formal explicit description of **classes** or **concepts** in a specific domain.
  - A class can have **subclasses** that represent concepts that are more specific than the superclass.
  - A class can have **instances**, i.e. actual representatives of classes.
  - Each concept can have **properties** describing various features, attributes and relationships of the concept.
  - **Facets:** Restrictions can be instantiated on properties (sometimes called **role restrictions**).

In Short

- GOLD is more than a taxonomy
  - It **incorporates** a taxonomy
- But it can also express a part/whole relationship, and restrictions on them
- Simply put, hierarchy is taxonomic, properties are part/whole

For Example



One Advantage of an Ontological Approach: Language Profiles

- Since it has the structure it does, an ontology will also allow you to define a subset of itself which typifies a particular language.
- It will also allow you to write software that can build very basic grammars of a language automatically.

Why is this?

- The ontology “knows” where individual pieces of data fit in a grammar.
- For example, suppose through the ontology we have defined the following for a language:
  - The language has six cases that go on nouns only.
  - It knows that nouns have three numbers in that language.

Why is this?

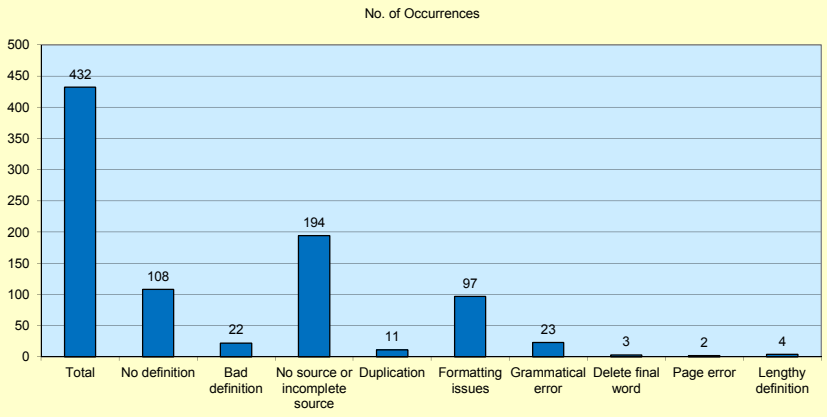
- Then…
  - It can build a paradigm for the noun.
- And it knows this has nothing to do with verbs…
- It can also write statements such as this “Nouns have six cases and three numbers, singular, dual and plural”
- So, you can tell it to automatically generate a very simple grammar based on this knowledge.

13

Refining GOLD

- December 2009 to July 2010
- GOLDComm team at LINGUIST List first checked every single existing definition & tracked any problems found
- Systematically solved the issues by adding/improving definitions & adding citations

Refining GOLD: Some statistics



Refining GOLD: Results

- Every concept in the ontology has a definition now (except OWL-specific ones)
- Unclear or insufficient definitions have been superseded
- All but ~52 definitions have citations now
- All new sources used go into the GOLD RDF bibliography for later import

Mapping the Content of GOLD to DCIF

- Relatively straightforward since the GOLD standard (as opposed to the web view) only contains concepts, their unique URI, their definitions and relations between the concepts

GOLD	DCIF
Concept Label	Data Element Name & Identifier
Concept URI	Source
Concept Definition	Definition (in Language Section)

Mapping the Structure of GOLD to DCIF

- Hierarchical/taxonomic relations from GOLD (parent-child-relationship, other relations need more work before adding them to DCIF)
- DCR not really intended to include complex relations
- Hierarchical relations from GOLD could only be partially retained

Mapping the Structure of GOLD to DCIF

- DCIF has <isA> element for simple data categories (DCs) and <value> elements in the Conceptual Domain section for complex/closed DCs
- DCs listed under the <value> element all have to be simple
- Complex/open DCs cannot have <value> element

Mapping the Structure of GOLD to DCIF

- Decision had to be made which concepts in GOLD hierarchy should become complex/closed DCs
- All children and lower would have to be simple DCs
- All parents and higher would become complex/open DCs and therefore not be linked to their children in any way

Mapping the Structure of GOLD to DCIF

Absolutive Antipassive Voice (Concept)

http://purl.org/linguistics/gold/AbsolutiveAntipassiveVoice

Thing
|_ Abstract
|_ Linguistic Property
|_ Morphosyntactic Property
|_ Voice Property
|_ Absolutive Antipassive Voice

Complex/open DCs

Complex/closed DC

Simple DCs

Definition:

A term used by some Mayanists to refer to an antipassive in which the patient or logical object is supressed or overtly absent. [Klaiman 1991: 232] Dayley states that the absolutive antipassive is used when the patient is unknown or irrelevant, or when the speaker does not wish to mention the patient, or to describe a transitive activity typically performed by some agent. The absolutive antipassive requires a non-specific implied patient, and no specific patient is ever semantically recoverable from the speech context. Only the absolutive antipassive has (as the name implies) absolutive function such that a transitive activity may be discussed without mention of the patient. [Dayley 1985: 345-350] This type of antipassive with an obligatorily unidentified object is also found in Mam, another Mayan language. [Cooreman 1994: 52-53]

Mapping the Structure of GOLD to DCIF

3326 **MorphosyntacticProperty** 1:0 private private closed gold-user private

MorphosyntacticProperty - 1:0

3. Conceptual Domain

Data Type	string
Profile	Morphosyntax
Value	/AbessiveCase/
Value	/AbilitativeModality/
Value	/AblativeCase/
Value	/AbsolutiveAntipassiveVoice/
Value	/AbsolutiveCase/
Value	/AccusativeCase/
Value	/ActionalFocus/
Value	/ActiveVoice/
Value	/AdessiveCase/
Value	/AgentDeletionPassiveVoice/
Value	/AllativeCase/
Value	/AnimateGender/
Value	/AntiCausativeVoice/
Value	/AntiPassiveVoice/
Value	/ApplicativeVoice/
Value	/ArabicNumeralGender/
Value	/AssumptiveEvidentiality/
Value	/AuditoryEvidentiality/
Value	/AugmentativeSize/
Value	/AversiveCase/
Value	/BenefactiveCase/
Value	/CaseProperty/
Value	/CausativeVoice/
Value	/ComitativeCase/

Child of VoiceProperty in GOLD

Child of MorphosyntacticProperty in GOLD

Summary & Outlook

- High level concepts not linked to their children and ancestry in DCR anymore
- Only by looking at <isA> elements of lower level concepts can the lineage to complex/closed DCs be traced
- There are plans to establish Relations Registry (using RDF/OWL) which could solve these issues

Thank you.

GOLD 2010:

http://linguistics-ontology.org/gold

GOLD Community website:

http://linguistics-ontology.org