*Workshop on the Feasibility of a Web-based Database of the Syntactic Structures of the World's Languages,*
NYU Linguistics, 9-10 Nov 2007

# Two foundational issues: the model of knowledge representation and the model of community

Gary Simons
*SIL International*

In reviewing the database proposal, I see two issues that are foundational to the design of the whole project. This presentation suggests two basic approaches to responding to each and argues that in each case the more complex approach is probably needed in order to truly achieve the stated objectives.

## 1. The model of knowledge representation

"The main purpose of the database will be to provide a tool for syntacticians, morphologists and semanticists doing cross-linguistic work which will allow them to explore the connections between various properties of the world's languages." The key term here is *property;* the formal characteristics of properties and their representation will determine what can be done with the database.

The simple approach:

- The system has an inventory of languages.
- The system has an inventory of properties and their possible values.
- The properties of languages are represented by assigning the appropriate values for each property.

Pros and cons:

- Easy to implement and use; but …
- Properties capture composite observations made by one analyst.
- The primitive facts of the grammar are not being represented.
- Users therefore cannot query to test hypotheses concerning the primitive facts or to form new composite observations involving the same primitives that underlie the properties that have been captured.

A more complex approach:

- The system has an inventory of languages.
- The system has an inventory of morphosyntactic resource types (e.g., syntactic category, feature, feature value) and the possible relationships between them.
- Morphosyntactic characteristics of the languages are formally described in terms of those resource types and relationships (e.g., a particular syntactic category in a particular language is inflected for a particular feature which has a particular set of possible values).
- Properties (in the sense of the simple approach) can be inferred by queries over those descriptions.

Cons and pros:

- Requires building consensus on the formal system of primitives.
- Requires learning a knowledge representation model in order to express morphosyntactic descriptions; but …
- A rich database of knowledge expressed in primitives is available to all researchers.
- Users can construct queries that build composite properties.
- Different researchers can explore different property sets for capturing composite observations over the same primitive knowledge.

RDF (Resource Description Framework) as a knowledge representation scheme:

- The KR scheme developed by the Semantic Web activity of the W3C (http://www.w3.org/RDF/).
- Represents each concept by a URI (Uniform Resource Identifier).
- Information is represented as a set of statements.
- Statement = a triple of < subject, predicate, object >, where:
    - The subject is a URI representing a *resource*.
    - The predicate is a URI representing a *property*.
    - The object may be another resource or it may be a *literal* value.
- A set of statements forms a directed graph.
- The basis for interoperation:
    - The graphs for individual descriptions can be merged into one large graph.
    - GOLD (http://www.linguistics-ontology.org/) already uses this approach.
    - The vocabulary of description can grow without requiring reprogramming.
- An RDF schema or OWL ontology is an RDF graph that formally defines the concepts (resource classes and properties) that are used in other RDF graphs.
    - *rdfs:Class* and *rdf:Property* are built-in resources.
    - *rdf:type* is a property to identify the class of which a particular resource is an instance.
    - *Rdfs:subclassOf* is a property that identifies one class as a subclass of another.
    - *rdfs:domain* and *rdfs:range* are properties that constrain the subjects and objects of properties.

Examples:

- See papers at: http://www.sil.org/~simonsg/by_subject.htm#Metaschemas
- Listing 1 gives the RDF schema for a "language profile" developed in EMELD project.
- Listing 2 gives the DTD for an XML representation of a language profile.
- Listing 3 is the XML source for parts of a language profile of Mutsun which was automatically translated to the RDF vocabulary by the metaschema processor.
- Sesame (an open-source RDF database with a Web interface) has a SeRQL query language:

*Languages with both number and case*

```
select distinct Language
from {DV} <gold:varietyOf> {Language},
     {DV} <gold:hasFeature> {} <rdf:type> {<gold:NumberFeature>},
     {DV} <gold:hasFeature> {} <rdf:type> {<gold:CaseFeature>}
```

*Languages with split ergativity*

```
select distinct Language
from {DV} gold:varietyOf {Language},
     {DV} gold: hasFeature {} gold:possibleValue {} rdf:type {gold:ErgativeCase},
     {DV} gold: hasFeature {} gold:possibleValue {} rdf:type {gold:NominativeCase}
```

## 2. The model of community

The project vision is to develop "an open and web-based community." What is the nature of that community: is it a community of end users or of projects and institutions? And what is the role of the database: is it to serve as the primary repository of user information or as a central aggregator of information held in a variety of locations? Answers to these questions will greatly affect the nature and probable impact of the project.

The simple approach:

- The project implements the database as the central *repository* of morphosyntactic knowledge.
- The project implements a web user interface that allows linguists to enter morphosyntactic knowledge.
- The project implements a web user interface that allows linguists to query the stored knowledge.

Pros and cons:

- Decision making and implementation is self-contained within the project.
- Provides a way for individual linguists who don't have a database to share what they know; but …
- Numerous projects that already have a wealth of knowledge in databases will not log in to re-enter what they know.
- This project will not allow users to explore all available cross-linguistic knowledge and its database will become one more silo competing for attention on the information landscape.

A more complex approach:

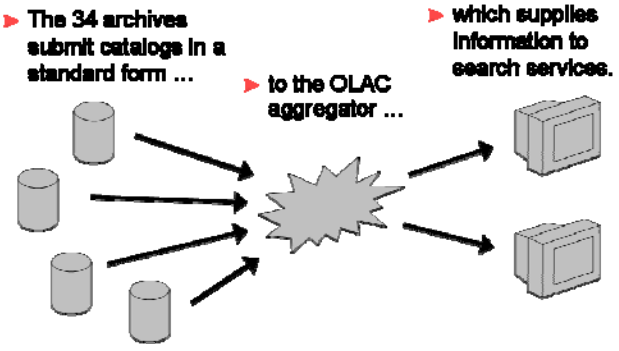
- The project implements the database as the central *aggregator* of morphosyntactic knowledge.
- The project defines a standardized knowledge representation form in which related projects can publish their morphosyntactic knowledge on their own web site.
- The project implements a web protocol for harvesting knowledge from participating projects.
- The project implements a web user interface that allows linguists to query the aggregated knowledge.

Cons and pros:

- Requires a great amount of cooperation and consensus building among already established projects.
- Project would also need to implement a participating repository service (or enlist a partner to do so) in order to establish a way for linguists who are not already part of a project to share; but ...
- The result is a community treasury of interoperable knowledge that is much greater than the sum of its parts.
- Cost is distributed across the community in a sustainable way since each participating institution supports its own participation.
- Primitive data items retain the identification of the contributing institution in their RDF resource URIs.

Examples:

- The Open Language Archives Community (http://www.language-archives.org/) is doing this with 34 institutions in the domain of metadata for language resource discovery.
- Operates on three standards: a knowledge representation standard (OLAC Metadata), a harvesting protocol standard (OLAC Repositories), and a community governance standard (OLAC Process).




- "Syntactic Structures of the World" could be built as a subcommunity within the OLAC framework, using OLAC infrastructure to register and find its members and their harvestable resources and to disseminate knowledge of what it offers to the wider linguistic community.
- See elements 7 and 8 of the vision in: http://linguistlist.org/tilr/papers/TILR%20Plenary.pdf

## Listing 1: RDF Schema for Language Profile

```
#
#  RDF Schema for Linguistic Description (in N3 notation)
#     These are additions to emeld.org/gold.owl (the instance-less version)
#  Gary Simons, SIL International
#  28 June 2004
#
        @prefix : <http://emeld.org/gold-ns#> .
        @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
        @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

# Described Variety

:DescribedVariety a rdfs:Class ;
        rdfs:subClassOf rdfs:Resource .

   :detailsOfVariety a rdf:Property ;
        rdfs:comment "Pinpoints the exact variety being described (as opposed to other varieties of the
                same language), such as the specific geographic region, age range, gender, social class" ;
        rdfs:domain :DescribedVariety ;
        rdfs:range rdfs:Literal .

   :describedBy a rdf:Property ;
        rdfs:comment "Names the person or persons responsible for this description. Ulitmately, it
                should be multiple references to a Person object." ;
        rdfs:domain :DescribedVariety ;
        rdfs:range rdfs:Literal .

   :varietyOf a rdf:Property ;
        rdfs:comment "The ISO639-3 languages of which this variety is an instance." ;
        rdfs:domain :DescribedVariety ;
        rdfs:range :ISO639Language .

:ISO639Language a rdfs:Class ;
        rdfs:comment "A distinct human language as recognized by the ISO 639-3 standard." ;
        rdfs:subClassOf rdfs:Resource .

# The language profile (i.e. grammar overview): Parts of Speech and Features

   :hasPOS a rdf:Property ;
        rdfs:comment "A parft of speech that is defined by the linguistic description." ;
        rdfs:domain :DescribedVariety ;
        rdfs:range :PartOfSpeech .

   :hasFeature a rdf:Property ;
        rdfs:comment "A morphosyntactic feature that is defined by the linguistic description." ;
        rdfs:domain :DescribedVariety ;
        rdfs:range :MorphosyntacticFeature .

 :PartOfSpeech a rdfs:Class ;
        rdfs:subClassOf rdfs:Resource .

   :inflectedFor a rdf:Property ;
        rdfs:comment "In this language, the part of speech is inflected for this feature." ;
        rdfs:domain :PartOfSpeech ;
        rdfs:range :MorphosyntacticFeature .

   :bearableFeature a rdf:Property ;
        rdfs:comment "An instance of this part of speech can inhereently bear this feature." ;
        rdfs:domain :PartOfSpeech ;
        rdfs:range :MorphosyntacticFeature .

 :MorphosyntacticFeature a rdfs:Class ;
        rdfs:subClassOf rdfs:Resource .

   :possibleValue a rdf:Property ;
        rdfs:comment "A possible value for the feature" ;
        rdfs:domain :MorphosyntacticFeature ;
        rdfs:range :FeatureValue .
```

```
 :FeatureValue a rdfs:Class ;
       rdfs:subClassOf rdfs:Resource .

# The morphosyntactic features like Number and Gender are defined in GOLD as subclasses of
# MorphosyntacticFeature.  The Number feature in a particular language is an instance
# of the Number class.  Similarly, FirstPerson is a subclass of FeatureValue.
```

## Listing 2: XML DTD for Language Profile

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v4.4 U (http://www.xmlspy.com) by Gary Simons (SIL International) -->
<!ELEMENT languageProfile (language, createdBy, date, source+, ontologyNamespace+, partsOfSpeech,
features)>
<!ELEMENT language (#PCDATA)>
<!ATTLIST language
  code CDATA #REQUIRED
>
<!ELEMENT createdBy (#PCDATA)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT source (#PCDATA)>
<!ELEMENT ontologyNamespace (#PCDATA)>
<!ATTLIST ontologyNamespace
  prefix CDATA #REQUIRED
>
<!ELEMENT partsOfSpeech (POS*)>
<!ELEMENT POS (label?, definition?, concept*, inflectedFor*, bearableFeature*)>
<!ATTLIST POS
  abbrev CDATA #IMPLIED
>
<!ELEMENT inflectedFor EMPTY>
<!ATTLIST inflectedFor
  target CDATA #REQUIRED
>
<!ELEMENT bearableFeature EMPTY>
<!ATTLIST bearableFeature
  target CDATA #REQUIRED
>
<!ELEMENT features (feature*)>
<!ELEMENT feature (label?, definition?, concept*, value*)>
<!ATTLIST feature
  abbrev CDATA #IMPLIED
>
<!--If <concept> is missing, then this value is mapped to an instance of the feature itself.-->
<!ELEMENT value (label?, definition?, concept*)>
<!ATTLIST value
  abbrev CDATA #IMPLIED
>
<!ELEMENT label (#PCDATA)>
<!ELEMENT definition (#PCDATA)>
<!ELEMENT concept (#PCDATA)>
<!ATTLIST concept
  relation (sameAs | similarTo) #IMPLIED
>
```

## Listing 3: XML Language Profile for Mutsun [css] (California)

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE languageProfile SYSTEM "LanguageProfile.dtd">
<languageProfile>
       <language code="CSS">Mutsun</language>
       <createdBy>Alexis Lanham</createdBy>
       <date>2004-07-02</date>
       <source/>
       <ontologyNamespace prefix="gold">http://emeld.org/gold-ns#</ontologyNamespace>
```

```
	<partsOfSpeech>
		<POS abbrev="n">
			<label>Noun</label>
			<concept>gold:Noun</concept>
			<inflectedFor target="case"/>
			<inflectedFor target="number"/>
			<inflectedFor target="sizeValue"/>
		</POS>
		<POS abbrev="v">
			<label>Verb</label>
			<concept>gold:Verb</concept>
			<inflectedFor target="Aspect"/>
			<inflectedFor target="Modality"/>
			<inflectedFor target="Person"/>
			<inflectedFor target="Voice"/>
			<inflectedFor target="Tense"/>
		</POS>
		<POS abbrev="Q">
			<label>Question word</label>
			<concept>gold:QuestionParticle</concept>
		</POS>
		...
	</partsOfSpeech>
	<features>
		<feature abbrev="Person">
			<label>Person</label>
			<concept>gold:PersonAttribute</concept>
			<value abbrev="3">
				<concept>gold:ThirdPerson</concept>
				<label>3rd</label>
			</value>
			<value abbrev="2">
				<concept>gold:SecondPerson</concept>
				<label>2nd</label>
			</value>
			<value abbrev="1">
				<concept>gold:FirstPerson</concept>
				<label>1st</label>
			</value>
		</feature>
		<feature abbrev="number">
			<label>Number</label>
			<concept>gold:NumberAttribute</concept>
			<value abbrev="PL">
				<concept>gold:Plural</concept>
				<label>Plural</label>
			</value>
			<value abbrev="SG">
				<concept>gold:Singular</concept>
				<label>Singular</label>
			</value>
		</feature>
		<feature abbrev="Aspect">
			<label>Aspect</label>
			<concept>gold:AspectAttribute</concept>
			<value abbrev="CONT">
				<concept>gold:ProgressiveAspect</concept>
				<label>Continuatuve</label>
			</value>
			<value abbrev="INT">
				<concept>gold:MorphosyntacticAttribute</concept>
				<label>Intensive</label>
			</value>
			<value abbrev="PERF">
				<concept>gold:PerfectiveAspect</concept>
				<label>Perfective</label>
			</value>
		</feature>
		...
	</features>
</languageProfile>
```