The semantics of markup: Mapping legacy markup schemas to a common semantics

Gary F. Simons
SIL International
7500 W. Camp Wisdom Road
Dallas TX 75236, USA
Gary_Simons@sil.org

William D. Lewis
Department of Linguistics
California State University, Fresno
5245 North Backer Avenue
Fresno CA 93740, USA
wlewis@csufresno.edu

Scott O. Farrar
Faculty of Linguistics and Literary Sciences
University of Bremen
Bibliothekstr. 1
D-28359 Bremen, Germany
farrar@uni-bremen.de

D. Terence Langendoen
Department of Linguistics
University of Arizona
P. O. Box 210028
Tucson AZ 85721, USA
langendt@u.arizona.edu

Brian Fitzsimons
Department of Linguistics
University of Arizona
P. O. Box 210028
Tucson AZ 85721, USA
fitzsimo@u.arizona.edu

Hector Gonzalez
Department of Linguistics
California State University, Fresno
5245 North Backer Avenue
Fresno CA 93740, USA
hexgonzo@csufresno.edu

Abstract

A method for mapping linguistic descriptions in plain XML into semantically rich RDF/OWL is outlined and demonstrated. Starting with Simons's (2003) original proof of concept of this method, we extend his Semantic Interpretation Language (SIL) for creating metaschemas to carry out the mapping, employ the General Ontology for Linguistic Description (GOLD) of Farrar and Langendoen (2003) as the target semantic schema, and make use of SeRQL, an RDF-aware search engine. This data migration effort is in keeping with the vision of a Semantic Web; it is part of an effort to build a 'community of practice' around semantically rich linguistic resources.

1 Introduction

Machine-readable structured linguistic documents (comparative word lists, lexicons, annotated texts, audio and audio-video recordings aligned with transcriptions (possibly annotated), grammatical descriptions, etc.) are being made available in a wide variety of formats on the Web. Until recently, the linguistics community has not been particularly concerned about the ease with which those structures can be accessed by other users, nor about the comparability of the structures that can be accessed. Now that community is beginning to realize that XML encoding provides relatively straightforward access to the intended structures and at the same time insures that the documents will continue be accessible for the foreseeable future.

However, XML encoding by itself does not insure comparability. To achieve that goal, the community must either adopt standards for encoding particular structures, or methods need to be developed for inter-

preting structures that are differently encoded. This paper reports on an effort to do the latter: to migrate XML documentation of linguistic structure to a semantically interoperable format. One of the most compelling reasons to do so is to enable *intelligent search*: the ability to query documents based on their semantics, rather than on strings of characters that may occur in them or on their document syntax. Facilitating intelligent searching is also one of the major goals of the Semantic Web. We are making the first steps towards a Semantic Web for linguistics by showing how to migrate a significant amount of language resources to a format that makes them semantically comparable.

2 Background

The work reported in this paper was carried out as part of the Electronic Metastructure for Endangered Language Data (EMELD) project [emeld.org] (NSF grant 0094934) and the Data-Driven Linguistic Ontology project (NSF grant 0411348). One of the objectives of the EMELD project is the "formulation and promulgation of best practice in linguistic markup of texts and lexicon." Underlying this objective is the goal of ensuring that the digital language documentation produced by linguists will be truly portable in the sense of Bird and Simons (2003): that it will transcend computer environments, scholarly communities, domains of application, and the passage of time. The project was undertaken on the basis of the following principles:

1. XML markup provides the best format for the interchange and archiving of endangered language description and documentation.
2. No single schema or set of schemas for XML markup can be imposed on all language resources.
3. The resources must nevertheless be comparable for searching, drawing inferences, etc.

Simons (2003) points out the conflict between the second and third principles, and describes the following set of actions for reconciling them.

1. Develop a community consensus on shared ontologies of linguistic concepts that can serve as the basis for interoperation.
2. Define the semantics of any particular markup schema by mapping its elements and attributes to the concepts in the shared ontology that they represent.
3. Map each individual language resource onto its (partial) semantic interpretation by applying the mapping of its markup schema.
4. Perform queries and other knowledge-based operations across resources over these semantic interpretations rather than the original XML documents.

The EMELD project has already begun work on the first of these action items, the creation of a sharable ontology for language documentation and description, a General Ontology for Linguistic Description (GOLD) [emeld.org/gold] (Farrar and Langendoen, 2003), which is intended to be grounded in a suitable upper ontology such as SUMO (Niles and Pease, 2001) or DOLCE (Masolo et al., 2002). GOLD is itself being written in OWL, the Ontology Web Language (McGuinness and van Harmelen, 2004), for use in Semantic Web applications. Simons (2003, 2004) also provides a 'proof of concept' for an implementation of the remaining three action items as follows.

1. Beginning with three dictionaries that used similar but distinct markup based on the Text Encoding Initiative (TEI) guidelines (Sperberg-McQueen and Burnard, 2002), Simons created mappings from their different markup schemas to a common semantics as defined by an RDF Schema (Brickley and Guha, 2004). Such a semantic schema provides a "formal definition ... of the concepts in a particular domain, including types of resources that exist, the properties that can relate pairs of resources, and the properties that can describe a single resource in terms of literal values" (Simons, 2004). This mapping he called a *metaschema*, a formal definition of how the elements and attributes of a markup schema are to be interpreted in terms of the concepts of the semantic schema. He called the 'language' for writing metaschemas (defined via an XML DTD) a Semantic Interpretation Language (SIL).

2. Simons performed the semantic interpretation operation in a two-step process using XSLT, first to create an interpreter for a particular metaschema and then to apply it against a source document to yield the RDF document (repository) that is its semantic interpretation.
3. Simons then loaded the RDF repositories into a Prolog system to create a merged database of RDF triples and used Prolog's inference engine to query the semantic interpretations.

Simons (2003) describes this implementation as providing a *semantics of markup*, rather than as devising yet another markup language for semantics. As such, it is in the spirit of efforts such as Sperberg-McQueen et al. (2000), who define the meaning of markup as the set of inferences licensed by it. However, their model does not provide for the general comparison of documents. It is also in the spirit of the proposal for a Linguistic Annotation Framework (LAF) under development by Working Group 1-1 of ISO TC 37 SC 4 [www.tc37sc4.org] (Ide and Romary, 2003; Ide, Romary and de la Clergerie, 2003), but differs from it in some significant ways. For example, our strategy does not require that the source annotations be mapped to an XML 'pivot format'. On the other hand, the LAF does not require that the source annotations be in XML to begin with. The 'data categories' of the LAF correspond to the concepts in GOLD; however the "creation of an ontology of annotation classes and types" is not yet part of the LAF (Ide, Romary and de la Clergerie 2003). Moreover, the LAF data model is confined to feature structures, whereas GOLD plans to offer feature structures as one of several data structuring alternatives. Finally, through its connection with an upper ontology, GOLD will also be related to the 'rest of the world', whereas the LAF ontology is apparently intended for linguistic structure only.

3 Goals of this paper

In this paper we extend Simons' proof of concept for the use of metaschemas in the following ways.

1. GOLD itself is used as the semantic schema.
2. SIL is extended to include the ability to map the content of designated elements and attributes in source documents to the semantic schema, not just the markup itself.
3. We devise metaschemas for lexicons that use distinct XML markup schemas: one of the lexicons that Simons (2003) originally used, for Sikaiana (Solomon Islands) with about 3000 entries; a Hopi (Arizona) dictionary with about 30,000 entries, for which Kenneth Hill's original encoding using a proprietary and no longer supported database program was converted to XML by Lewis and Gonzalez; and a Potawatomi (Great Lakes region, US and Canada) lexicon being created by Laura Buszard-Welcher using the EMELD FIELD tool.
4. The Prolog query engine is replaced by SeRQL, an SQL-like query language for Sesame, an RDF database program (Broekstra, Kampman and van Harmelen 2002; *User Guide for Sesame* 2004). It is our intention to couple Sesame with an inference engine that reads OWL documents, such as Racer (Haarslev and Moller 2001).

In carrying out the migration of such language resources to the Semantic Web, we are guided by the principle of preserving the original analyses as much as possible. At the same time, since the migrated resources are to be rendered mutually interoperable and transparent to the tools that are designed to work over them, the migration process has the potential to greatly increase the precision of the original analyses, to reveal inconsistencies in them, and ultimately to result in enriched resources. For example, the comparison of two descriptions of the same language that has been made possible by migration could reveal errors in one or the other. Similarly, a single resource could be checked for consistency with accumulated linguistic knowledge represented in an ontology. The migration process thus provides two sources of new knowledge. First is the knowledge brought in from the document interpretation process itself, i.e. by the linguist, not necessarily the one who performed the original analysis. Second when the migrated documents are added to the knowledge base, new inferences can be automatically generated based on the general knowledge of linguistics captured in the ontology. The type of new knowledge generated is however constrained, for example, by the type of search to be done over the resulting knowledge base (see section 6).

However the migration process can also skew or misinterpret the intentions underlying the original documentation. To minimize this risk, the migration tools should be as non-intrusive as possible. Even so, some steps are necessary to add structure where structure is lacking in the original XML documentation and to interpret the meaning of the original elements where their meanings are undefined or unclear. For the ontology the implication is that theory-laden concepts either should be avoided or less encumbered alternatives should be made available.

4 GOLD

An important guiding principle used in the construction of GOLD is to distinguish between those concepts that represent the content of linguistic data and those that pertain to the structuring of those data (cf. Ide and Romary 2003 who also distinguish between data content and data structure). A particular entry in a lexicon, for example, is a data structure used to organize lexical data in a particular fashion. Entries usually contain actual data instances, e.g., the Hopi word *nahalayvi'yma* or its phonological properties. The process of data migration is made much easier if a separation between data and data structure is upheld in the semantic schema.

4.1 Data content

Linguistic data content includes linguistic expressions, the physical manifestations of language, also known as 'morphs', or simply 'forms', which may be written, spoken or signed. In GOLD, written linguistic expressions are represented as ORTHOGRAPHICEXPRESSION with the subclasses ORTHOGRAPHICPART, ORTHOGRAPHICWORD, and ORTHOGRAPHICSENTENCE. These are defined as special types of strings. In order to analyze linguistic data further, abstract counterparts of linguistic expressions are proposed called LINGUISTICUNIT. The abstract units are the main objects of interest in formal linguistics. In some theories, the various subclasses of LINGUISTICUNIT correspond to 'morphemes', 'constituents', or 'constructions'. No assumptions are made about whether these have any mental significance, e.g. whether they are underlying forms. The class hierarchy for LINGUISTICUNIT is presented in Farrar, Lewis and Langendoen (2002), and can be viewed in GOLD using Protégé 2.0 [protege.stanford.edu].

The LINGUISTICUNIT hierarchy is organized according to how its components are realized as forms, and not according to their formal linguistic features, which are theory specific. So, for example, LEXICALUNIT is simply a formal unit that can appear in isolation in its realized form, and not necessarily something that can be a constituent of larger syntactic constructions. The methodology leaves open the question of whether, for example, a SUBLEXICALUNIT can also be a phrasal constituent, as appears to be the case with CLITIC. Yet another alternative would be to organize LINGUISTICUNIT according to semantic features, e.g., a SUBLEXICALUNIT would be something which usually represents a grammaticized notion. But, since this varies from language to language, a different taxonomy would be needed for every type of language encountered. To sum up, adhering to strictly formal features necessitates theory-specific taxonomies, while adhering to semantic features leads to language-specific taxonomies. Instead a neutral approach is taken in which LINGUISTICUNIT is organized according to how instances are realized as linguistic expressions.

ORTHOGRAPHICEXPRESSION is related to LINGUISTICUNIT by the predicate REALIZES. The particular sort of LINGUISTICUNIT is further defined according to what kinds of attributes it can take. So, a MORPHOSYNTACTICUNIT has attributes of the sort MORPHOSYNTACTICATTRIBUTE. Instances of particular attributes are PASTTENSE, SINGULARNUMBER, and PROGRESSIVEASPECT. The class of attributes pertaining to linguistic units parallels other kinds of non-linguistic attributes such as SHAPEATTRIBUTE and PHYSICALSTATE.

There are several varieties of attributes which linguists find useful for language description, including phonological and semantic features. Semantic attributes contrast with morphosyntactic attributes in that the former correspond to the notional characteristics of linguistic form that have some manifestation in the grammar.

4.2 Data structures

A linguistic data structure is defined as an abstract information container which provides a way to package elements of linguistic data. The two main types of data structures contained in GOLD at the moment are LEXICALITEM and FEATURESTRUCTURE. Our characterization of LEXICALITEM extends that of Bell and Bird (2000). At a minimum, a LEXICALITEM should contain an instance of LEXICALUNIT or of SUBLEXICALUNIT. Special relations are given in GOLD which pertain only to data structures, e.g., HASLEXICALUNIT relates a LEXICALITEM to a LEXICALUNIT. Instances of LEXICALITEM typically include glosses either in the same language in the case of a monolingual lexicon, or in some other language in the case of a bilingual lexicon. Glosses are simply instances of ORTHOGRAPHICEXPRESSION related to the entry via the relation GLOSS. Entries relate to one another via relations such as SYNONYMOF and ANTONYMOF.

If a LEXICALITEM contains extensive morphological information, we may represent this in the form of a FEATURESTRUCTURE. The FEATURESTRUCTURE class is part of a more extensive set of data structures known as a FEATURESYSTEM (Langendoen and Simons, 1995; Maxwell, Simons and Hayashi, 2002). A FEATURESPECIFICATION is a data structure that contains a subclass and an instance of MORPHOSYNTACTICATTRIBUTE (i.e. an ordered pair), for example, [TENSE: PASTTENSE]. The implementation of the FEATURESYSTEM construct allows for recursive FEATURESPECIFICATIONs in which, for example, a subclass of MORPHOSYNTACTICATTRIBUTE is paired with an instance of FEATURESTRUCTURE.

One criticism that could be raised against the inclusion of data structures in a semantic resource such as GOLD is that they are superfluous. Why not simply leave it up to the source markup to describe the elements of data structure, e.g., in the form of an XML Schema? This is certainly a reasonable criticism, since excluding data structures from GOLD would make the ontological modelling process much simpler. However, they are included because we envision that subsequent applications will need to be able to reason, not only about the data itself, but also about how it is structured. For example, it might be necessary to compare elements of a LEXICALITEM to that of FEATURESTRUCTURE. This is actually an essential step in achieving the vision of the Semantic Web, namely, constraining the source data in such a way as to preserve structure where structure is defined and to enrich structure where structure is left unspecified.

5 Semantic Interpretation Language

The Semantic Interpretation Language (SIL) was originally created to define the meaning of the elements and attributes declared in an XML markup schema, as well as the relationships between them. An SIL *metaschema* is an XML document that formally maps the elements and attributes of an XML encoded resource to concepts in an OWL ontology or an RDF Schema. Furthermore, the metaschema formally interprets the original markup *structure* by declaring what the dominance and linking relations in the XML document structure represent. For example, consider the extract from the Hopi lexicon shown in Figure 1.

The dominance relation between the elements `<MSI>` (for 'morphosyntactic information') and `<POS>` (for 'part of speech') in the original XML is implicitly something like 'has'. This can be made more explicit by mapping it to HASMORPHOSYNTACTICPROPERTY, a formally defined relation in the ontology. This relation is formally defined in the ontology by specifying its *signature*, i.e. what kinds of arguments it can take. Thus, a better defined, more exact, relationship between elements of markup is achieved.

```
<Lexeme id="L3">
 <Headword>naheva</Headword>
 <MSI>
  <POS>
   <Feature name = "type">vt
         </Feature>
  </POS>
```

Figure 1. Extract from Hopi Lexicon

SIL has been extended to formalize the resolution of *content* in addition to markup. For example, the semantics of the gram `vt` in the XML structure `<POS>vt</POS>` can be specified via a mapping to the ontology as an instance of VERBTRANSITIVE, in addition to defining the semantics of the `POS` element itself.

An SIL metaschema, as described in detail in Simons (2004), is an XML document built from metaschema *directives*, which are essentially processing instructions expressed as XML elements. Directives like `resource`, `property`, `literal` and `translate` generate elements of the resulting semantic interpretation. Part of the SIL DTD is shown in Figure 2.

```
<!ELEMENT metaschema (namespace+, (interpret | ignore)+)>
<!ELEMENT namespace (#PCDATA)>
<!ATTLIST namespace prefix CDATA #REQUIRED>
<!ELEMENT interpret (resource | translate | property | literal)*>
<!ATTLIST interpret markup CDATA #REQUIRED>
<!ELEMENT resource (property | translate | literal | embed)*>
<!ATTLIST resource concept CDATA #REQUIRED>
<!ELEMENT property (resource | resourceRef | embed)>
<!ATTLIST property concept CDATA #REQUIRED>
<!ELEMENT translate EMPTY>
<!ATTLIST translate concept CDATA #REQUIRED mapping CDATA #REQUIRED>
```

Figure 2. SIL DTD fragment

The `interpret` directive performs the primary mapping function from markup elements of the input resource to the enriched output, as demonstrated in Figure 3. The tag `<form>` is interpreted as a LINGUISTICFORM, specifically as an ORTHOGRAPHICREPRESENTATION of that form.

Input document:

```
<form>ahali</form>
```

Metaschema directive:

```
<interpret markup="form">
    <property concept = "gold:form">
    <resource concept = "gold:LinguisticForm">
      <literal concept = "gold:orthographicRepresentation"/>
    </resource>
  </property>
</interpret>
```

Interpretation (output):

```
<gold:form>
   <gold:LinguisticForm>
     <gold:orthographicRepresentation>ahali
     </gold:orthographicRepresentation>
   </gold:LinguisticForm>
 </gold:form>
```

Figure 3. Example interpretation of an element

Of primary importance to the interpretation of content is the *translate* directive, as shown in Figure 4. In this example, the tag `<Feature name="type">`, embedded within `<POS>`, is interpreted as referencing a morphosyntactic property, the value of which is content interpretable by the *terminology set* identified by the reference `Hopi/Hopi_pos_mapping.xml`. A terminology set contains a simple

mapping between terms used in the source document and the names of the equivalent concepts in the ontology. SIL can handle both one-to-one terminology mappings (e.g., mapping from the tag `vt` to the concept VERBTRANSITIVE) as well as one-to-many mappings (e.g. mapping from `1sg` to a property bundle of FIRSTPERSON and SINGULARNUMBER).

Input document:

```
<POS>
   <Feature name = "type">vt </Feature>
</POS>
```

Metaschema directive:

```
<interpret markup = "POS/ Feature[@name='type']">
 <translate concept = "gold:property" mapping =
   "Hopi/Hopi_pos_mapping.xml"/>
</interpret>
```

Interpretation (output):

```
<gold:property rdf:resource = "emeld.org/gold#VerbTransitive"/>
```

Figure 4. Example interpretation of content

SIL is designed to allow interoperability between resources by mapping the different structures and content of markup in the source documents onto the same set of ontological concepts. This is demonstrated by comparing the transformed output for Hopi shown in Figure 4 with the transformed output for Sikaiana in Figure 5. Note that the inputs are different but the outputs are the same.

Input document:

```
<pos>Verbt</pos>
```

Metaschema directive:

```
<interpret markup="pos">
<translate concept = "gold:property" mapping =
   "SKY/SKY_pos_mapping.xml"/>
</interpret>
```

Interpretation (output):

```
<gold:property rdf:resource = "emeld.org/gold#VerbTransitive"/>
```

Figure 5. Transformed Sikaiana <pos>

The SIL only guarantees interoperability when comparable semantic resources are employed in the mapping. If an entire group relies on a common semantic schema, e.g. GOLD, a ‘community of practice’ is formed. This in turn facilitates *intelligent search* across converted resources.

Currently, writing an SIL metaschema is done entirely by hand. We are in the process, however, of developing two tools to automate the process. The first tool will allow the user to define the relationship between the terminology used within a resource with relevant GOLD concepts. The second tool will define the structural mapping relationship between the resource and a given meta-structure. The first tool, named Alchemy, presents the user with a drag-and-drop interface in which the user defines the terms used within her resource by associating them with one or more GOLD concepts. The relationship between any given term and relevant GOLD concepts can be complex, with one-to-one or one-to-many relationships being allowed, and the relationships themselves can be of any of a number of types: SameAs, KindOf, etc. We are in the process of building this tool, embedded within an systems developer toolkit accompanying GOLD.

The second as of yet unnamed tool is still in the early design stages. This tool will allow the user to first define the type of resource she is converting (lexicon, interlinear text, grammar, etc.), and will then lead

her through a series of questions that define the structure by associating it with a meta-type definition for the particular resource type. The tool will require a precise and well-defined 'semantics of linguistic structure', a conceptual space of linguistic structural types that will be included in GOLD, but is still in the process of being defined. The final output of this tool, in association with an Alchemy-defined terminology set, will be an SIL metaschema.

6 Querying Resources

In this section, we discuss the general issue of searching over linguistic descriptions on the Web, and the current state of our effort to do so using SeRQL (see section 3 item 4) over the RDF repositories for Sikaiana, Hopi and Potawatomi generated by the metaschemas from their XML-encoded lexicons.

6.1 Dimensions of search over linguistic descriptions

As mentioned in section 1 above, one of the most compelling reasons to migrate XML documentation to a semantically interoperable format is to enable *intelligent search*. For the linguistics community, we envision several parameters of search over semantically interoperable linguistic documentation. Search may be performed according to:

- level of analysis (phonetic, morphosyntactic, discourse)
- typological dimension (including language type)
- intent of search (for exploring some particular language, or for language comparison)
- kind of results desired (which data structure to return)

Search also varies according to degree of difficulty, that is, whether search requires the assistance of an inferencing engine or not. *Direct search* is defined as search over explicitly represented data, i.e. instance data in the knowledge space. This includes the simple string matching of conventional search engines. But since the search will be carried out using the enriched RDF framework, direct search is not limited to string matching in the original XML. An example of direct search is to find all data that includes a reference to instances of some grammatical category (e.g., PASTTENSE). Boolean searching with direct search is also possible, e.g., searching for cases of portmanteau morphemes, expressed in our framework as two or more MORPHOSYNTACTICATTRIBUTEs associated with some LINGUISTICUNIT.

Indirect search goes beyond direct search by making use of inferences based on the structuring of the concepts in an ontology. For example the concept of PLURALNUMBER means 'two or more', the concept of DUALNUMBER means 'exactly two', and the concept of MULTALNUMBER means 'three or more'. A direct search for PLURALNUMBER will miss those instances represented as DUALNUMBER and MULTALNUMBER, whereas an indirect search will find them.

6.2 Some SeRQL queries

In Figure 6, we give the SeRQL query (omitting `using namespace`) for the orthographic forms for all the lexical items specified as having the GOLD concept PROGRESSIVEASPECT in the three lexicons. This query returned 1135 results, all from Hopi.

```
select distinct R
from {LI} <gold:meaning> {} <gold:grammar> {} <gold:property>
   {<gold:ProgressiveAspect>},
{LI} <gold:form> {} <gold:orthographicRepresentation> {R}
```

Figure 6. SeRQL query for PROGRESSIVEASPECT forms

Next, the query in Figure 7 returns all the grammatical properties of lexical items categorized as NOUNs in each of the lexicons. There were 21 results from Hopi, 3 from Sikaiana and 6 from Potawatomi; an example for each language is given in Figure 8. The fact that certain items categorized as NOUNs in Sikaiana are also categorized as VERBs indicates that those items have both classifications. In Figure 9, we give the SeRQL query for all such items; 61 results were obtained.

```
select distinct P, LC
from {LI} <gold:meaning> {} <gold:grammar> {MSI} <gold:property>
  {<gold:Noun>};
<gold:property> {P},
{LI} <gold:languageCode> {LC}
where P != <gold:Noun>
```

Figure 7. SeRQL query for attributes of NOUNs

Hopi: AUGMENTATIVE
Sikaiana: VERB
Potawatomi: INANIMATE

Figure 8. Sample results of query in Figure 7

```
select distinct LI
from {LI} <gold:meaning> {} <gold:grammar> {} <gold:property>
   {<gold:Noun>};
   <gold:property> {<gold:Verb>}
```

Figure 9. SeRQL query for all lexical items marked as both NOUN and VERB

Finally in Figure 10, we give a query used to find the parts of speech that are common to entries in the Hopi and Sikaiana lexicons. Four results were returned, NOUN, VERB, ADJECTIVE and NUMERAL.

```
select distinct P
   from {LI} <gold:meaning> {} <gold:grammar> {} <gold:property> {P},
   {LI2} <gold:meaning> {} <gold:grammar> {} <gold:property> {P},
   {LI} <gold:languageCode> {LC},
   {LI2} <gold:languageCode> {LC2}
where LC = "HOP" AND LC2 = "SKY"
```

Figure 10. SeRQL query for common parts of speech in two lexicons

More complex queries that take advantage of the structure of the ontology are also possible, for example to find all the verbs in the lexicons regardless of whether they have been tagged as transitive verbs, intransitive verbs, or simply as verbs. With further development of the method described here, much more elaborate queries over much larger linguistic data repositories will be possible. This result, we hope, will encourage much more widespread distribution of language resources on the Web and the creation of a large community of practice that uses those resources for research, teaching, and language revitalization efforts.

References

J. Bell and S. Bird. 2000. *A preliminary study of the structure of lexicon entries*. In "Workshop on Web-Based Language Documentation and Description", Philadelphia. [www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html]

S. Bird and G. F. Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3):557-582.

D. Brickley and R. V. Guha. 2004. *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation 10 February 2004, World Wide Web Consortium. [www.w3.org/TR/rdf-schema]

J. Broekstra, A. Kampman and F. van Harmelen. 2002. *Sesame: A generic architecture for storing and querying RDF and RDF schema*. In "Proceedings of the First International Semantic Web Conference", I. Horrocks & J. Hendler, ed., pages 54-68, Springer-Verlag, Berlin.

S. O. Farrar and D. T. Langendoen. 2003. A linguistic ontology for the Semantic Web. *Glot International* 7(3):97-100.

S. O. Farrar, W. D. Lewis and D. T. Langendoen. 2002. *An ontology for linguistic annotation*. In "Semantic Web Meets Language Resources: Papers from the AAAI Workshop", N. Ide & C. Welty, ed., pages 11-16, AAAI Press, Menlo Park, CA.

V. Haarslev and R. Moller. 2001. *Description of the RACER system and its applications*. In "Proceedings of the Description Logics Workshop DL2001", pages 132-142, Stanford, CA.

N. Ide and L. Romary. 2003. *Outline of the international standard Linguistic Annotation Framework.* In "Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right", pages 1-5, Sapporo. [www.cs.vassar.edu/~ide/papers/acl2003-ws-laf.pdf]

N. Ide, L. Romary and E. de la Clergerie. 2003. *International standard for a Linguistic Annotation Framework*. In "Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology", Edmonton.
[www.cs.vassar.edu/~ide/papers/ide-romary-clergerie.pdf]

D. T. Langendoen and G. F. Simons. 1995. A rationale for the Text Encoding Initiative recommendations for feature-structure markup. *Computers and the Humanities* 29:191-205.

C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari and L. Schneider. 2002. *WonderWeb deliverable D17 version 2.0*. In "The WonderWeb Library of Foundational Ontologies and the DOLCE ontology." [www.loa-cnr.it/Papers/WonderWebD17V2.0.pdf]

M. Maxwell, G. F. Simons and L. Hayashi. 2002. *A morphological glossing assistant*. In "Proceedings of the International Workshop on Resources and Tools in Field Linguistics", Las Palmas, Spain.
[www.mpi.nl/lrec/papers/lrec-pap-25-MorphologicalGlossingAssistant.pdf]

D. L. McGuinness and F. van Harmelen, ed. 2004. *OWL Web Ontology Language overview*.
[www.w3.org/TR/2004/REC-owl-features-20040210]

I. Niles and A. Pease. 2001. *Toward a standard upper ontology*. In "Proceedings of the 2nd International conference on Formal Ontology in Information Systems", Ogunquit, ME.
[projects.teknowledge.com/HPKB/Publications/FOIS.pdf]

G. F. Simons. 2003. *Developing a metaschema language to support interoperation among XML resources with different markup schemas.* Paper presented at the ACH/ALLC conference, Athens, GA.
[www.sil.org/~simonsg/metaschema/ACH%202003.pdf]

G. F. Simons. 2004. *A metaschema language for the semantic interpretation of XML markup in documents*. Technical report, SIL, Dallas. [www.sil.org/~simonsg/metaschema/sil.htm]

C. M. Sperberg-McQueen and L. Burnard, eds. 2002. *TEI P4: Guidelines for electronic text encoding and interchange, XML version*, Text Encoding Initiative Consortium, Oxford etc. [www.tei-c.org/P4X]

C. M. Sperberg-McQueen, C. Huitfeldt, and A. Renear. 2000. Meaning and interpretation of markup. *Markup Languages: Theory and Practice* 2:215-234.

User Guide for Sesame. 2004. [www.openrdf.org/publications/users/index.html]