The GOLD Community of Practice

An Infrastructure for Linguistic Data on the Web

Scott Farrar
Universität Bremen

William D. Lewis
University of Washington

Abstract. The GOLD Community of Practice is proposed as a model for managing on-line linguistic data. The key components of the model include the linguistic data resources themselves and those focused on the knowledge derived from data. Data resources include the ever-increasing amount of linguistic field data and other descriptive language resources being migrated to the Web. The knowledge resources capture generalizations about the data and are anchored in the General Ontology for Linguistic Description, or 'GOLD'. It is argued that such a model is in the spirit of the vision for a Semantic Web and, thus, provides a concrete methodology for rendering highly divergent resources interoperable. Furthermore, a methodology is given for creating specific communities of practice within the overall scientific domain of linguistics. Finally, a number of applications are proposed including those aimed at knowledge acquisition and those aimed at putting the knowledge to use.

Keywords: best practice, markup, linguistics, ontology, Semantic Web, smart search, interlinear glossed text

1. Introduction

While there is no available statistic, the amount of electronically available linguistic field data seems to be increasing at a phenomenal rate. A simple Web query for even the most obscure language can yield scholarly papers containing richly annotated data, entire websites dedicated to the description of the language or language family, or even posted field notes with sound and video files. While the situation opens up enormous opportunity for automated empirical research, we will argue that such a rapid increase in the number of Web resources motivates the need for community consensus concerning the quality control of data, agreement in terms of encoding and markup formats, and according to common tools and supporting resources.

In this paper, we discuss a general Web architecture whereby community consensus can be achieved. The formation of such a community addresses many problems created by the explosion of electronically available data by: (1) fostering of diverse sub-communities united towards a common scientific goal; (2) developing a scalable migration strategy from data to knowledge; and (3) providing a semantically

interoperable format suitable for intelligent search over very large-scale data stores. Central to the community is the codification of the knowledge of linguistics. We take advantage of one such effort, the **General Ontology for Linguistic Description** (GOLD) (Farrar and Langendoen, 2003; Farrar, in Press). Based on GOLD, then, we present a detailed model for a community of practice centered around linguistic data on the Web, which we call the **GOLD Community of Practice**. The general idea of the GOLD Community is to provide an architecture (websites, resources, and tools) such that each of its components makes use of GOLD, and one that is suitable to the needs and technical expertise of the average linguist. The GOLD Community of Practice provides linguistics with a way to take advantage of recently standardized technologies such as XML, RDF, and OWL. Much in the spirit of the Semantic Web (Berners-Lee et al., 2001), the Community provides the means whereby linguists can use diverse terminology, yet arrive at a consensus through what the markup elements, or terms, mean and, thus, achieve true data interoperability.

In Section 2 we give the relevant background concerning the nature of linguistic data and the various challenges that such data pose for creating and maintaining a community of practice. In Section 3 we describe the individual components that make up the GOLD Community. In Section 4 we describe various applications built in support of the community of practice. Finally, in Section 5 we provide a summary and a discussion concerning the broader impacts of this research.

2. Background

This section attempts to focus the discussion by providing the relevant backround concerning linguistic data. We focus in particular on what makes linguistic data both challenging and well suited for incorporation into a knowledge-based model. Then, we turn to a description of 'best-practice' in terms of data encoding and markup. Finally, we give an overview of the relevant aspects of the Semantic Web.

2.1. The Nature of Linguistic Data

That descriptive linguistic data are already available on the Web – in fact, in large amounts – means that linguistics has the opportunity to utilize the Web as the primary means of data access and management, if not for the entire field, at least for particular sub-disciplines. The process of creating a useable framework is, however, much more difficult than just collecting relevant URLs or creating a specialized linguistics

search engine. The situation is due largely to the fact that linguistic data is heterogeneous. For example, the terminology used to describe data can be based on specific theoretical assumptions that are not likely to be relevant for, and not likely to be mappable to, other data resources. Nevertheless, we can make some key generalizations that reveal the nature of linguistic data, and thus suggest a treatment within a unified framework.

For expository purposes, consider the data instance given in (1). This instance is typical of that found in the descriptive linguistics literature and illustrates some key features of linguistic data.

(1) Keq=apc sesolahki=te mihqitahas-iyin ehcuwi-monuhmon-s?
what=again suddenly=Emph rem-2Conj IC.must-buy.2Conj-DubPret
What else did you suddenly remember you had to buy? (Bruening, 2001)

Line 1 contains actual data content, in other words, linguistic expressions such as *Keq*. Usually the product of linguistic field research, data content is any element that is essentially unanalyzed. Though in practice, data content will contain implicit analysis, e.g., in the form of phonemic segmentation. Lines 2 and 3 contain elements of data analysis, which is anything that is not data content. Examples of data analysis include a morphological breakdown of words in a language (as in line 2), a translation (as in line 3), a syntactic description of some sentence, a comparison of two lexicons, etc. Analysis shows up in documents in a form that Bird and Liberman (2000) have called 'annotation'. Referring to example (1), we distinguish terms used to label elements of analysis from the elements themselves, or those entities posited to exist by the linguist. That is, what are actually given in line 2 of (1) are the (abbreviated) terms themselves. In order to make sense of such terms, we need to know their intended meaning, something that is often missing from the analysis of linguistic data. Scholarly papers, dictionaries, and grammars about a particular language will often append an informal terminology set as a guide. But even so, the terms used in linguistic analyses may remain largely ambiguous (Langendoen et al., 2002). For example, the term *NOM* could be used to label either NominativeCase or Nominalizer. On the other hand, different terms can often have the same intended semantics, especially across different analyses. Finally, the example itself is given as **interlinear glossed text** (IGT), a very common linguistic data structure, an organizational device for grouping together data content and analysis for some display, theoretical, or computational purpose. From this simple example then, we may discern three aspects of the data: the data content itself, the components of data analysis, and components of the data structure. We argue that most, if not all, linguistic data have these

three components, and that this bears directly on being able to treat all kinds of data in a unified framework.

2.2. Best-Practice Encoding of Linguistic Data

With various key aspects of data in focus, we turn now to some of the issues related to its encoding and markup. As our point of departure we refer to the results of the E-MELD project [emeld.org] whose primary aim has been the promulgation of the **best-practice principles**. E-MELD builds on the work of the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 2002) and also the work of Bird and Simons (2003). At a minimum these principles require the consistent use of Unicode (The-Unicode-Consortium, 2000) and XML (Bray et al., 2004) to encode and mark up data content. XML contrasts, for example, with less structured formats including HTML and text documents and with proprietary file formats. In terms of the elements of data analysis, E-MELD and Bird and Simons (2003) recommend mapping all terminology to a machine-readable semantic resource that defines them. It was also proposed in Farrar et al. (2002) to map all terms in a domain-specific ontology for linguistics.

Finally, we note an alternative for dealing with diverse terminology in linguistics: the development scientific standards. The Linguistic Annotation Framework (LAF), for example, is under development by ISO TC37 SC4 Working Group 1-1, (Ide and Romary, 2003) and (Ide et al., 2003). As part of the broader ISO effort (TC37) devoted to standardization of 'Terminology and Other Language Resources', the SC4 subcommittee is in general devoted to 'Language Resource Management', that is, "to prepare international standards and guidelines for effective language resource management in applications in the multilingual information society" (Ide and Romary, 2003, p. 1). As our experience with a diverse group of linguists has shown, arriving at a common, accepted standard is nearly impossible, though some standards such as the ISO 639 or the Ethnologue three-letter language codes (see http://www.ethnologue.com/.) seem to be gaining acceptance. Instead of relying on the universal acceptance, and use, of such standards, we opt to give *any* term used in the descriptive and analytic markup of data a machine-readable semantics.

2.3. The Vision of the Semantic Web

Even if such best-practice methods were fully implemented and accessbile on-line, it would still be difficult to achieve effective machine processability without more sophisticated markup. The Web as it is

currently known is an environment designed for and accessible to humans. In recent years, however, there has been an ever-growing focus on creating Web content that can be processed by machines. A key requirement for such a task is a way to represent what things on the Web mean. This emphasis on meaning is most clearly articulated in what has become known as the 'Vision of the Semantic Web' (Berners-Lee et al., 2001). The success of *the* Semantic Web has perhaps been overestimated as if one day the Web as we know it will suddenly be switched off and a new a semantically enriched Web put in its place. This is not the case. While there has been a leap in the number of available Semantic Web technologies (e.g., ontology languages), there exist no hard and fast solutions for creating a Semantic Web. However, it is our claim that, at least for individual sciences, the Semantic Web is achievable when approached from the bottom up. That is, it is the responsibility of particular scientific communities such as linguistics or chemistry to first create a Semantic Web for their own disciplines. Only then could it ever be expected that they will merge and thus help to achieve the loftier vision.

3. Components of the GOLD Community

The GOLD Community of Practice consists of two types of components: those centered around linguistic data, and those concerning general linguistic knowledge. The data-centric components compose the empirical part of the model, one in which data is represented both in its raw form and as semi- or fully analyzed data. The knowledge centric components capture the collective knowledge of the field, that which is ultimately grounded in data. The aim of this section is to bring out this difference and to explain how the components compose a unified whole. This discussion will set the stage for Section 4 where we present a detailed discussion of how knowledge can be derived from data within the model.

3.1. Data-centric Components

As the GOLD Community is primarily designed to take advantage of the rapidly growing descriptive material on the world's languages, the core of the Community is the data upon which it is built. Ideally, the GOLD Community would be based on those resources in a best-practice format, which minimally requires the consistent use of Unicode and XML. However, since the move towards such best-practice formats is a relatively slow process – especially considering the

long tradition of display-centric data representation – the Community should also accommodate for so-called **legacy** resources, essentially those display-centric, proprietary resources which are not in XML. The following section describes each type of data resources while attempting to emphasize the need for best-practice.

3.1.1. *Best-practice Resources*

Based on Bird and Simons (2003) and the discussion in Section 2.2, we adopt the general concept of 'best practice' for linguistic data. We refer to a collection of linguistic data that conforms to such a recommendation as a **best practice resource**. In terms of the GOLD Community of Practice model, the most important requirements for such a resource involve its encoding and markup. That is, the encoding should be Unicode, while the markup scheme should be XML accompanied by a DTD or Schema. More structured and semantically oriented formats are available, e.g., RDF(S) or OWL, but these formats are more appropriate for implementing the knowledge components (to be discussed in Section 3.2). Thus, for a general data format maintainable by the average working linguist, we argue that a basic Unicode/XML encoding is sufficient, and even desirable over the richer formats. The main reason is that the XML data model is, in general, easier for linguists to apply, not to mention that a broad variety of software is available for manipulating XML documents. We argue, in fact, that XML encourages linguists to follow best-practice recommendations, because it does not involve a major time commitment for mastery. (For a further discussion of the merits of XML, see Bird and Simons, 2003.)

In terms of particular XML structures, we encourage the use of DTDs or Schemas already developed or recommended by the E-MELD project. A key characteristic of such resources is that – as discussed in Section 2.2 – they are focused more on description and less on display. The main reason for preferring descriptive over display-centric XML is that adequate display can always be derived from well described content. Descriptive content is in a sense more fundamental than display, since the same content can be rendered in a number of different ways. Consider, for example, that whereas traditional print dictionaries are ordered according to alphabetic or similar orthographically-based criteria, electronic dictionaries can be presented according to rhyming patterns, root morpheme, or even frequency. The point is that display follows from description and not vice versa, and that data should ultimately be maintained in descriptive format. The rendering of descriptive data into a display-centric format is best considered as a separate application that can be built around the GOLD Community of

Practice. In fact, we argue that rendering is one of the key applications that will ensure the Community's success. If data is renderable in a multitude of different display formats, then many different groups can access the data in ways that make sense for them. This is particularly important when considering a dilemma often encountered in linguistic field work – namely, how to balance the needs of the scientific community with the needs of the speaker community. Linguistic research demands a display organized according to analysis, while the speaker community could be better served with a display organized, for example, to benefit language learners.

3.1.2. *Termsets*

One of the primary goals of the GOLD Community is to draw on empirical data in order to augment the general knowledge of the field. This requires mapping individual data sets to knowledge-based components. Even with well designed best-practice resources in place, the mapping process would be a daunting task and in most cases be beyond manual effort. Instead, the mapping will be at best semi-automated. But any hope of automation requires something beyond best-practice. One of the primary reasons is the inconsistent or ambiguous use of markup terminology. Whereas many linguists already use terminology commonly accepted in their subfield, the wider audience across the entire field may not recognize it. Some markup elements could be considered as standard or at least near-standard, e.g., *3PL* or *ACC*. But without a theoretical context, it could be impossible to determine the meaning of terms, e.g., *NOM*, *CL*, *PST*, etc. What is needed is an explicit definition of what markup elements mean. Therefore, we suggest the use of **termsets** to supplement any best-practice data resources.

We define a termset as a mapping from a set of markup elements T, used in a data resource, to a set of classes or instances C from the GOLD ontology.

Definition 1 (Termset). A termset is the tuple $\langle T, C \rangle$, where:

- T is a set of markup elements and C is a set of classes or instances from GOLD;
- For each $t \in T$, there is zero or more $c \in C$ such that t 'denotes' c;
- If there is more than one c for a given t, i.e., $\{c_1, c_2, \ldots, c_n\}$, then interpret the set as the union $c_1 \cup c_2 \cup c_n$.

The definition states that, where possible, markup elements should be mapped to a single concept in an ontology. Although it is possible for a markup element, even within a limited community, to represent more than one concept (for example, *NOM* representing either 'nominative case' or 'nominalizer'), we require that only elements with a conjunctive

meaning, e.g., *3SG* or *1PL*, be used in this manner. Note that the definition does not preclude the use of identical terms in two or more disjoint data resources, where the two do not share the same termset. Thus, if *NOM* in resource R_1 is mapped to NominativeCase, then *NOM* in R_2 can be mapped to something besides NominativeCase. Finally, it is allowable for multiple terms to represent the same element in GOLD. Here is a snippet of an XML implementation of a termset for a description of the language Maasai that conforms to the definition.

```
<terms>
    <term ID="NOM">
       <concept>gold:NominativeCase</concept>
       <comment>Nominative Case subjects</comment>
    </term>
    <term ID="3SG">
       <concept>gold:ThirdPerson</concept>
       <concept>gold:SingularNumber</concept>
       <comment>Third person and singular
       number.
       </comment>
    </term>
    <term ID="CL">
       <comment>CL marks nominal classifiers.
       Referent unknown.
       </comment>
    </term>
    ...
 </terms>
```

Note that the term *CL* in the example termset is not defined according to the ontology, and thus does not conform to the requirement that each term be defined. We still consider such a resource a well-formed termset, however, the data described by *CL* would not be accessible through search and other tools that use the ontology. We allow such flexibility in a termset in cases, for example, where the ontology contains no appropriate concepts for a given term, or it is unclear what the referent is. We expect that the number of gaps in the ontology will decrease as more and more data is considered and the ontology is augmented.

A termset is intended to be used as input to an automated processor for migrating the data to an interoperable format, which will be discussed separately as a supporting application in Section 4. Mapping to an ontology, other than simply providing semantic grounding, facilitates such applications as "smart" or concept-based search. For example, a query for the concept Singular would return data described

with markup elements such as *SG*, as well as *1SG*, *2SG*, *SING*, etc. Furthermore, the use of termsets encourage the formation of communities of practice based on shared terminology. In this way, linguists can use or at least relate their own terms to ones that have been previously recognized within a community.

Finally, we also encourage the use of terminologies as developed by such standardization efforts as ISO TC37 SC4 Working Group 1-1, (Ide and Romary, 2003) and (Ide et al., 2003). One aim of ISO TC37 SC4 is to develop 'data category registries' (Romary, 2003). The advantage of using such registries is that they reflect quasi-standard uses of terminology by experts in particular subfields, especially with regards to the markup of data from majority languages. The structure of the proposed data category registries is precisely in line with the GOLD Community of Practice and can also be useful for the markup of lesser studied languages.

3.1.3. *Descriptive Profiles*

Termsets are, in a sense, snap-shots of grammars and are easily created. At a minimum they indicate what categories a grammar contains and can be used to achieve some degree of interoperability among disparate data resources. They do not, however, provide the means to say anything definitive about grammatical systems, such as "these are *all* the cases of a language" or "aspect is marked only on modal verbs". This is precisely the kind of knowledge that work in descriptive linguistics is intended to capture. Therefore, for greater interoperability, it is necessary to go beyond simple termsets and to formulate a resource with potentially much more structure. This resource should capture some portion of the grammar via a **grammar fragment**. A grammar fragment is defined as a formalization of some portion of a language's grammatical system.

Definition 2 (Grammar Fragment). A grammar fragment is a tuple $\langle C, L_{DS} \rangle$, where:

- C is a set of linguistic concepts;
- L_{DS} is a set of formal data structures;
- Each $c \in C$ is contained in some $l_{DS} \in L_{DS}$.

The definition is rather broad stating that a grammar fragment can include any kind of useful, systematic grammatical information, e.g., the possible morphophonemic combinations of a language or co-occurrence constraints on morphosyntactic features (such as the work discussed in Section 2.2). The only requirement is that the data structures be defined in GOLD or an extension thereof. A grammar fragment is

just that, a *fragment*, because the knowledge of a language's structure, function, etc. will almost always be incomplete, especially in cases of preliminary field data reports.

Termsets and fragments, then, are different, but interrelated parts of a language description. With their definitions in place, we introduce the next type of data-centric resource meant to bring the termsets and grammar fragments together, namely, the **descriptive profile**. Inspired by work on the FIELD (Aristar, 2003), a tool for automatically producing profiles of lexicons, we propose that a descriptive profile include a termset and one or more grammar fragments.

Definition 3 (Descriptive Profile). A descriptive profile is the tuple $\langle T_s, G \rangle$, where:

- T_s is a termset;
- G is a grammar fragment;
- The data expressed in each $g \in G$ is expressed using T_s.

Here is a descriptive profile for Georgian that meets these minimal requirements.

```
<profile>
   <termset>
      <term>
      ...
   </termset>
   <gram>
      <feature>gold:Case
         <value>gold:NominativeCase</value>
         <value>gold:AccusativeCase</value>
         <value>gold:DativeCase</value>
         <value>gold:ErgativeCase</value>
         <value>gold:GenitiveCase</value>
      </feature>
      <feature>gold:Aspect
         <value>gold:PerfectiveAspect</value>
         <value>gold:ImperfectiveAspect</value>
      </feature>
   </gram>
   ...
</profile>
```

First, there is a termset indicating what the markup elements mean. The termset is followed by a grammar fragment listing of all the morphological cases of the language. Grammatical fragments are useful

within the GOLD Community, because they facilitate the derivation of *new* knowledge from best-practice resources. While it is possible to conclude that the entire case inventory of Georgian is given in the list of terms in the example above, it would not be a sound inference since this information is not explicitly given. A grammatical fragment provides the means to state explicitly that, for example, the given case values are exhaustive.

3.1.4. *Legacy resources*

Most linguistic data on the Web at the moment reside in resources that do not conform to best practice, what we refer to as **legacy resources**. Legacy resources are either in semi- or unstructured formats, such as HTML or plain text documents, or proprietary formats, including PDF and various word-processing formats which cannot be read in the absence of special software (e.g., Microsoft Word or Excel). Many such formats, especially the proprietary ones, are notoriously difficult to process automatically (Bird and Simons, 2003). Generally speaking, the linguistic data structures used in these formats are **display oriented**, designed to accommodate the needs of displaying the data in a human readable format. Display-oriented devices have a long tradition in the field of linguistics and play a key role in the human readability of data in scholarly publications. Display-oriented formats contrast with those required for linguistic theories and computational systems. This latter type are more compact and in general are more explicity structured.

A major advantage of the GOLD Community of Practice is that it also accommodates legacy resources, primarily because, once a mechanism for migration or access is in place, the availability of legacy data would be of immense importance to the field and broaden the scope of the GOLD Community of Practice. But *in lieu* of costly software to directly migrate legacy to best-practice resources, a kind of short-cut can be achieved by migrating only the most important, descriptively relevant aspects. This migration would involve constructing a descriptive profile for the resource, which would encode a terminology mapping and any relevant grammar fragments. A summary of the various data-centric components is given graphically in Figure 1.

3.2. Knowledge-centric components

Having introduced the fundamental data-centric components, we now turn to a description of the components concerned with knowledge. The main role of the knowledge-centric components is to represent the knowledge that is captured explicitly or implicitly by the data. On the one hand, the knowledge-centric components capture the general,

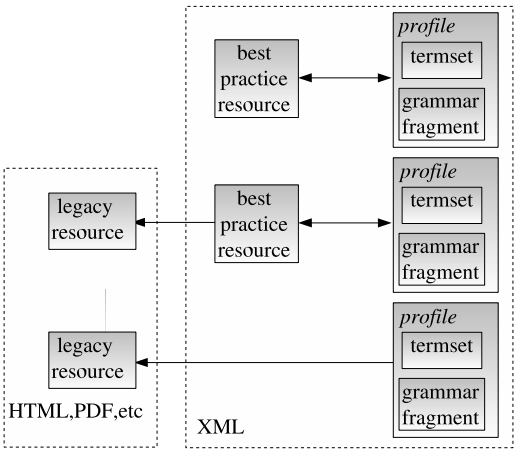

Figure 1. Data-Centric Components of the GOLD Community

canonical knowledge of the field. On the other, they represent the knowledge that is verifiable in empirical data. With this move from data to knowledge, we take particular inspiration from the field of knowledge engineering and the recent work in applied formal ontology, in particular how these fields can be used to model specific scientific domains. One of the key problems that the GOLD Community addresses is the control and separation of various knowledge components. As will be discussed in the following section, the design provides a means to separate (1) general linguistic knowledge from (2) the knowledge of particular languages and from (3) knowledge that pertains only to specific sub-communities of practice. Furthermore, the design allows for relating the linguistic knowledge of the GOLD Community to an upper ontology. In short, we provide a realization of the vision of the Semantic Web for descriptive linguistics.

3.2.1. *The General Ontology for Linguistic Description*
The most central knowledge component of the GOLD Community of Practice is the **General Ontology for Linguistic Description** (GOLD) (Farrar and Langendoen, 2003; Farrar, in Press). Thus far, in the description of data-centric components, we have focused on individual linguistic descriptions that capture the knowledge common to a particular theory or specific to an analysis. In contrast to this type of knowledge is that which can be considered as canonical, or at least widely accepted – the general knowledge of the field that is usually possessed by a well trained linguist. This includes knowledge that po-

tentially forms the basis of any theoretical framework. In particular, GOLD captures the fundamentals of descriptive linguistics. Examples of such knowledge are 'a verb is a part of speech', 'gender can be semantically grounded', or 'linguistic expressions realize morphemes'.

The list above shows knowledge of a generic sort, that which is typically represented in the ontologies of expert and knowledge-based systems. The modeling choices in GOLD are described elsewhere (e.g., Farrar, in Press); therefore here, we only mention a some key aspects of its implementation relevant for the GOLD Community of Practice. For instance, we note that whereas GOLD *could* be used to represent linguistic universals, e.g., in the sense of Greenberg (1966), we choose not to include them. Instead, the intention is to include the necessary meta-knowledge from which inferences regarding universals could be drawn. That is, the derivation of implied universals could be given as a potential application on top of the GOLD Community of Practice.

From a practical knowledge-engineering standpoint, it is difficult to separate general linguistic knowledge from that which pertains to specific languages. After all, the scientific knowledge of Hopi, English, and Ancient Greek is all part of the canon of linguistics. For example, that Hopi has an ImperfectiveAspect or that English and Greek both have a PastTense constitute linguistic knowledge; but, this kind of knowledge can be differentiated from the general knowledge listed above, as it is only relevant for specific languages. A similar issue is differentiating between theory-specific knowledge and that which pertains to the entire field. We do not claim that GOLD is completely theory independent, but we do claim that its categories are at least applicable to a diverse set of linguistic theories. Therefore, we reserve GOLD for capturing the most general sorts of linguistic knowledge, and propose other resources to capture the more theory- or language-specific knowledge. In the next section we describe these resources along with a general solution to the problem of how to keep such sorts of knowledge separate.

Finally, we note that GOLD is grounded in an **upper ontology**, or one that provides the basic tools for constructing any ontology including the ontology meta-language itself (e.g., subclass, instance, set theory), a theory of basic mereology, a theory of roles, a theory of action, etc. For this we are experimenting with the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001; Pease and Niles, 2002) and also the Descriptive Ontology for Language and Cognitive Engineering (DOLCE) (Masolo et al., 2003). We will not describe these ontologies here, though interested readers are referred to Farrar and Bateman (2004) for a discussion and evaluation of various upper ontologies.

3.2.2. *Community of Practice Extensions*
When linguists assume language-specific or theory-specific knowledge, they are essentially identifying their research with a particular sub-community of practice within linguistics. Sub-communities are readily distinguishable by the terminology employed in data annotation. If it were only a question of terminology, then many-to-many, or even simple term mappings could be constructed as part of the data components to achieve a high degree of interoperability. But linguists not only employ different surface terminologies, they actual conceptualize their disciplines in divergent, and often, incompatible ways.

To capture explicitly the relationship between sub-community knowledge and GOLD, we include a knowledge resource called a **Community of Practice Extension** (COPE). A COPE is an extension of GOLD, a sub-ontology, that inherits all or a portion of GOLD's conceptualization depending on the specific requirements of the sub-community.

Definition 4 (COPE). A COPE is the tuple $\langle C, I \rangle$, where:

- C is a set of classes and I is a set of individuals in the COPE;
- $\forall x.C(x)\exists y.C(x) \rightarrow C_g(y)$, such that C_g is a class in GOLD;
- $\forall x.inst(x, C)\exists y.inst(x, y) \rightarrow C_g(y)$;
- If C_i and C_j are COPEs, then $C_i \cup C_j$ is also a COPE.

To explain the definition, a COPE is first of all a set of classes C and instances I. Second, all classes must be subsumed by some GOLD class C_g or by a class from some other COPE, that is in turn subsumed by a GOLD class. Third, all individuals in a COPE must be direct or indirect instances of GOLD classes, allowing for indirect instantiation via another GOLD anchored COPE. The requirements for what can be a COPE are not very stringent. In fact, a new COPE can be constructed entirely of classes and individuals from one or more other COPEs, as shown in the final statement. More commonly, however, a COPE will be constructed by using only a few "recycled" classes and individuals. Consider the scenario where a COPE C_i is being created for a given language family and there already exists a general COPE C_j for the grammatical category of aspect. Then, some of C_i's members will be included in C_j.

Thus, we envision several types of COPEs. First, consider that for a description of languages such as Swahili and related languages, a focused and an extensive knowledge pertaining to a Bantu NounClass is required. Such a COPE would facilitate the definition of a Bantu ProtoNounClass and could be shared across the Bantu community. Second, the concepts of some linguistics subdomains could be con-

structed and maintained relatively separately; consider for instance, the sub-discipline of phonetics. Phonetics fundamentals, e.g., SEGMENT or PITCH, could be captured in a single COPE and shared across a wide community. Thirdly, a COPE could be constructed to capture concepts that are particular to a given data type. A lexicon COPE, for example, would utilize concepts such as LEXEME, HEADWORD, or SUBENTRY. Finally, consider the diversity of conceptualization found in, for example, Minimalism and Systemic Functional Grammar. Whereas some of the basic conceptualization is shared, e.g., the existence of LINGUISTICFEATURE, a concept like MERGE would only be relevant for Minimalism, and a concept such as IDEATIONALUNIT would only pertain to Systematic Functional Grammar. That is, concepts native to particular theoretical perspectives are best kept separated into different COPEs.

Thus, the use of COPEs furthers the modular design of the model. A summary of the various knowledge components is given in Figure 2. From the figure, the main relation of subsumption (SB), which holds

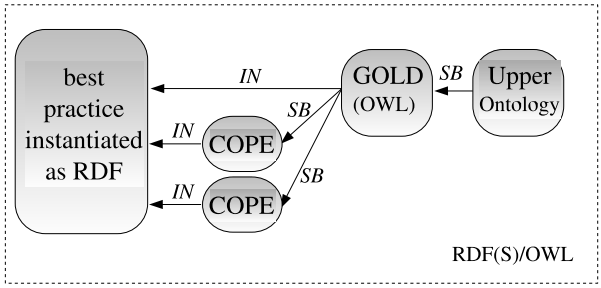

Figure 2. Knowledge-centric components of the GOLD Community

between classes, links the various ontologies and sub-ontologies to one another. The various COPEs are subsumed by GOLD, and GOLD is in turn subsumed by the upper ontology. But there is also the relation of instantiation (IN) that holds between individuals and classes. The individuals are the actual elements of data content, analysis and data structure that have been migrated from the best-practice XML documents. Note that an element of data, analysis, or structure need not instantiate a particular COPE, but can directly instantiate a GOLD concept (or upper ontology concept), as shown by the direct IN link to GOLD.

With the knowledge component fully explicated, we can now give a more precise definition of a **sub-community of practice** within the community of practice: it is the consistent application of a particular COPE that may, but does not require, the use of the compatible ter-

minology. With COPEs, communities have the ability to maintain the knowledge central to their community in discrete, manageable packets. This provides at least two benefits. First, from a knowledge engineering perspective, individual COPEs can be mined to add missing knowledge to GOLD. Consider the scenario where GOLD is lacking a particular fundamental tense category, and where there exist several detailed COPEs that encompass a description of tense. If the tense category is really fundamental, then it should show up in numerous COPEs. The common knowledge captured by the different COPEs can then be migrated to GOLD, and formally structured according to the rest of general linguistic knowledge, thus obviating the need for future COPEs to "re-create" the knowledge. Second, the separation makes sense because it provides a simple method of control over what types of knowledge are considered in queries. That is, if a user wants to exclude certain language-specific knowledge from their queries (if the analysis is in question, or irrelevant), then by having this knowledge separated into various COPEs, the exclusion can be done in the query component by simply deselecting a particular data source.

4. Applications Based on the GOLD Community of Practice

The Community will provide a dynamic 'workplace' for making the most of linguistic data, achievable using various Web applications built around the static infrastructure. The applications can be divided into those used for the *acquisition* of knowledge by the GOLD Community and those concerned with the *application* of that knowledge.

Transforming data into knowledge is not a simple process, but requires an advanced Web application. For example, various terminologies used in best-practice resources first need to be rendered transparent and compatible with one another by mapping them onto a set of descriptive profiles. As an example of the migration process itself, we draw on the work of Simons (2003) and Simons (2004) in which the Semantic Interpretation Language (SIL) is developed to transform semi-structured data (in XML) to highly-structured knowledge (in RDF(S)/OWL). The SIL is a generalized framework implemented using XML and the Extensible Stylesheet Language (XSL) (W3C, 2001) that formally maps the elements and attributes of best practice XML resources to a common 'semantic schema'. The schema can be in the form of an RDF Schema or OWL ontology. The strength of the SIL, then, is that it provides the means to manipulate the original XML at both the syntactic and the semantic level. Central to the SIL is the notion of a **metaschema** (Simons, 2003). The metaschema is a document consist-

ing of a set of directives in the SIL language that instructs the processor how to interpret the original best practice markup elements in terms of the concepts of a semantic schema. Furthermore, the metaschema formally interprets the original markup structure by declaring what the dominance and linking relations in the XML document structure represent. We have demonstrated in Simons et al. (2004a) and Simons et al. (2004b) that the migration process can be successfully implemented in a scalable, systematic fashion.

Recall from Section 3.1.4, however, that most of the linguistic data currently on the Web is contained in legacy resources. The ubiquity of data that exist in legacy formats argues for a mechanism of extracting data from such resources, or minimally providing systematic access to them. Also recall that certain kinds of semi-structured data are common in linguistic discourse and are often encapsulated in documents encoded in proprietary file formats. There is some potential for the automated extraction or migration of display data types from proprietary file formats to richer data formats, such as XML. If the XML format conforms to best-practice recommendations, then a migration to knowledge-centric components is readily achieveable (as shown in Simons et al. (2004a) and Simons et al. (2004b)). And, if done on a large enough scale, the full migration could help to ensure the acceptance of the GOLD Community itself. For more discussion of the automated conversion of legacy resources into particular types of semi-structured data formats, see Lewis (2003) and Simons et al. (2004a). For an implementation using migrated IGT see the Online Database of Interlinear Text (ODIN) (http://www.csufresno.edu/odin).

We finally come to the question of how the GOLD Community of Practice can be put to use. The first and perhaps most important application that the GOLD Community will provide is **ontology-driven search** over massive amounts of disparate data. There are essentially two types of ontology-driven searches envisioned within the GOLD Community: *concept* searches and *intelligent* searches. The former makes minimal use of the ontology whereby users specify a concept as the search parameter. The query engine then searches across a semantically normalized database to find all instances of data that instantiate that concept. This differs significantly from simple string-matching searches that are typical in database and Web environments. For example, in a typical string search on the Web, searching for "PST" might return instances of data containing 'past tense' morphemes, but it is equally likely to return documents concerning 'Pacific Standard Time'! On the other hand, a more intelligent concept search for Subject would return data that are marked for all of the following: *Subject*, *SUBJ*, *NOM*, and *ERG* (ErgativeCase). Such a query might also

return ABS (AbsolutiveCase) if the query engine is able to discern the relationship of the noun so marked and the type of verb: AbsolutiveCase marks the subject of intransitive verbs and the object of transitive verbs. Such a search might be an instance of *intelligent search* since an inference might need to be made with respect to the relationship between the verb and noun if that relationship is not explicitly marked in the description.

An example of concept search used in Simons et al. (2004b) is: "List language data for all languages where one word encodes both PastTense and SecondPerson." The query returned an instance of data (see Example 1) from the Passamaquoddy IGT data set, the only instance that satisfied the condition. Note that the *-s* morpheme instantiates the Preterite, a form of the past tense, the morpheme *monuhmon* marks *2Conj*, a form of SecondPerson, and that both morphemes are in the same word. An intelligent search infers meaning from a query, such that the full power of the ontology and the knowledge base is tapped to find data and analyses that may not have been explicitly asked for, but are relevant to the query nonetheless. For example, if we pose the query "List all the objects of verbs in Yaqui", the query engine could use the ontology to infer that by 'objects' we mean 'nouns' (or 'noun phrases') since nouns are typically objects of verbs. It could also infer that nouns that are objects of verbs must be marked for a case appropriate to object position. In nominative/accusative languages like Yaqui, such a noun would be marked for AccusativeCase. Thus, the search actually performed is "List all instances of Noun marked for AccusativeCase in Yaqui".

5. Summary and Discussion

We have presented a model for a community of practice centered around linguistic data on the Web and GOLD, an ontology for linguistic description. The model was designed based on the nature of linguistic data. It was inspired from recent efforts to establish best-practice encoding and markup schemes, especially that suggested by Bird and Simons (2003) and the E-MELD project. To implement the model, we have drawn on numerous Web technologies including XML, RDF(S), and OWL. We have shown how such an implementation is an instantiation of the vision of the Semantic Web for the linguistics domain. We have described the individual components of the model, divided into those centered around data and those centered around knowledge about that data. We have shown that the primary benefit of the model is that community control over individual data resources is maintained, yet a

high degree of interoperability is achieved among disparate resources. We have noted that profiles are a tangible artifact useful for the creation of specific communities of practice, centered around a consensus about what terms mean.

Acknowledgements

A special thanks goes to Terry Langendoen for his support of our research project from the beginning. The idea to construct an ontology for linguistics was conceived by the authors during their work on the Electronic Metastructure for Endangered Language Data (E-MELD) project [emeld.org] (NSF ITR-0094934). For this endeavor, we gratefully acknowledge the support of the E-MELD PIs and associates, especially Gary Simons, Helen Aristar-Dry and Anthony Aristar. We acknowledge the comments of the members of the "GOLD summit" held in November, 2004 in Fresno, CA, including Jeff Good, Baden Hughes, Laura Buszard-Welcher, Brian Fitzsimons, and Ruby Basham. Finally, we gratefully acknowledge the NSF-funded Data-Driven Linguistic Ontology Development project (BCE-0411348) which supported the authors during the writing of this manuscript.

References

Aristar, A.: 2003, 'FIELD.Lex'. Technical report, presented at Workshop on Digitizing and Annotating Texts and Field Recordings, LSA Institute.

Berners-Lee, T., J. Hendler, and O. Lassila: 2001, 'The Semantic Web'. *Scientific American*.

Bird, S. and M. Liberman: 2000, 'A formal framework for linguistic annotation'. Technical Report MS-CIS-99-01, Computer and Information Science, University of Pennsylvania.

Bird, S. and G. Simons: 2003, 'Seven dimensions of portability for language documentation and description'. *Language* **79**.

Bray, T., J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau: 2004, 'Extensible Markup Language (XML) 1.0 (Third Edition)'. Technical report, World Wide Web Consortium (W3C).

Bruening, B.: 2001, 'Syntax at the Edge: Cross-Clausal Phenomena and the Syntax of Passamaquoddy'. Ph.D. thesis, MIT.

Farrar, S.: in press, 'Using 'Ontolinguistics' for language description'. In: A. Schalley and D. Zaefferer (eds.): *Ontolinguistics: How ontological status shapes the linguistic coding of concepts*. Berlin: Mouton de Gruyter.

Farrar, S. and J. Bateman: 2004, 'General ontology baseline'. SFB/TR8 internal report I1-[OntoSpace]: D1, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany.

Farrar, S. and D. T. Langendoen: 2003, 'A linguistic ontology for the Semantic Web'. *GLOT International* **7**(3), 97–100.

Farrar, S., W. D. Lewis, and D. T. Langendoen: 2002, 'An ontology for linguistic annotation'. In: *Semantic Web Meets Language Resources: Papers from the AAAI Workshop, Technical Report WS-02-16*. Menlo Park, CA, pp. 11–19.

Greenberg, J.: 1966, *Language Universals*. The Hague: Mouton.

Ide, N. 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*.

Ide, N., L. Romary, and E. D. la Clergerie: 2003, 'International Standard for a Linguistic Annotation Framework'. In: *Proceedings of NAACL'03 Workshop of Software Engineering and Architecture of Language Technology Systems*.

Langendoen, D. T., S. Farrar, and W. D. Lewis: 2002, 'Bridging the markup gap: smart search engines for language researchers'. In: *Proceedings of the International Workshop on Resources and Tools in Field Linguistics*. Las Palmas, Gran Canaria, Spain.

Lewis, W. D.: 2003, 'Mining and Migrating Interlinear Glossed Text'. Technical report, Workshop on Digitizing and Annotating Texts and Field Recordings, LSA Institute. Available at http://emeld.org/workshop/2003/papers03.html.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari: 2003, 'Ontologies library (final)'. WonderWeb Deliverable D18, ISTC-CNR, Padova, Italy.

Niles, I. and A. Pease: 2001, 'Toward a standard upper ontology'. In: C. Welty and B. Smith (eds.): *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Ogunquit, Maine.

Pease, A. and I. Niles: 2002, 'IEEE Standard Upper Ontology: A Progress Report'. *Knowledge Engineering Review: Special Issue on Ontologies and Agents* **17**.

Romary, L.: 2003, 'Implementing a data category registry within ISO TC37– Technical note contributing to a future WD for ISO 12620-1'. Technical Report SC36N0581, International Standards Organization.

Simons, G.: 2004, 'A metaschema language for the semantic interpretation of XML markup in documents'. Technical report, SIL.

Simons, G. F.: 2003, 'Developing a metaschema language to support interoperation among XML resources with different markup schemas'. In: *Proceedings of the ACH/ALLC conference*. Athens, GA.

Simons, G. F., B. Fitzsimons, D. T. Langendoen, W. D. Lewis, S. O. Farrar, A. Lanham, R. Basham, and H. Gonzalez: 2004a, 'A model for interoperability: XML documents as an RDF database'. In: *Proceedings of the EMELD Workshop on Databases*. Detroit, MI.

Simons, G. F., W. D. Lewis, S. O. Farrar, D. T. Langendoen, B. Fitzsimons, and H. Gonzalez: 2004b, 'The semantics of markup: Mapping legacy markup schemas to a common semantics'. In: *Proceedings of the 4th workshop on NLP and XML (NLPXML-2004)*. Barcelona, Spain, pp. 25–32. held in cooperation with ACL-04.

Sperberg-McQueen, C. M. and L. Burnard: 2002, *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlottesville, and Bergen: Text Encoding Initiative Consortium.

The-Unicode-Consortium: 2000, *The Unicode Standard, Version 3.1.1, defined by: The Unicode Standard, Version 3.0*. Reading, MA: Addison-Wesley.

W3C: 2001, 'Extensible Stylesheet Language (XSL) Version 1.0'. Recommendation, W3C. Available at http://www.w3.org/TR/2001/REC-xsl-20011015/.