The GOLD Community of Practice

An Infrastructure for Linguistic Data on the Web

Scott Farrar
University of Arizona

William D. Lewis
University of Washington and California State University Fresno

Abstract. The GOLD Community of Practice is proposed as a model for linking on-line linguistic data to an ontology. 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 (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 semantically interoperable. The focus of this work, then, is not on annotation at the syntactic level, but rather on how annotated Web resources can be linked to an ontology. Furthermore, a methodology is given for creating specific communities of practice within the overall Web infrastructure for linguistics. Finally, ontology-driven search is discussed as a key application of the proposed model.

Keywords: descriptive linguistics, best practice, markup, ontology, Semantic Web, smart search

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 (family), or even field notes with sound and video files. While the situation opens up enormous opportunities for automated empirical research, we will argue that such a rapid increase in the number of Web resources motivates the need for community consensus to take advantage of the Web as the primary resource for accessing data. Community consensus concerns several factors, including the quality control of data, an agreement on encoding and markup standards, and the use of common tools and supporting resources. There have been several research projects aimed at developing standards and frameworks for handling data in fields such as NLP and language engineering. While these projects certainly offer many lessons, there are no comparable off-the-shelf frameworks for the scaleable management of descriptive linguistic resources. Furthermore, there has been an ever-growing focus on creating Web resources that can be processed by machines. Broadly, this has become known as the "Vision of the Semantic Web" (Berners-Lee et al., 2001). While there has been a surge 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 disciplines, be it 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 of total content integration.

In this paper, we discuss a general Web architecture whereby community consensus can be achieved focusing *not* on the format or structure of annotation, but rather on the intended meaning of markup. The formation of such a community addresses many problems created by the explosion of electronically available data by: (1) fostering diverse subcommunities 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). Using GOLD, then, we present a detailed model for a community of practice centered around linguistic data on the Web, called the **GOLD Community of Practice**, or *GOLDComm*. GOLDComm is an architecture (tools and data/knowledge resources) 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. GOLDComm provides linguistics with a way to take advantage of recent markup technologies (XML, RDF(S), and OWL). Much in the spirit of the Semantic Web, GOLDComm provides the means whereby linguists can use diverse terminology, yet arrive at a consensus through what the markup elements mean, thus achieving true content interoperability over diverse data through the use of ontologies. This work is not intended to replace or compete with existing annotation frameworks, but to supplement them by suggesting a mechanism for migration from richly annotated resources to semantically grounded ones.

In Section 2 we give 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 components that make up GOLDComm. In Section 4 we review several projects related to the current one and give relevant comparisons. Finally, in Section 5 we discuss a how the model can be used to achieve intelligent, ontology-driven search.

2. Background

2.1. The Nature of Linguistic Data

Descriptive linguistic data are already available on the Web in large amounts, meaning that linguists have the opportunity to utilize the Web as the primary means of data access and management. 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, both raw and annotated, are 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.

To illustrate some key features of linguistic data, consider the Passamaquoddy data in (1), which is typical of that found in the descriptive linguistics literature.

(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, i.e., linguistic expressions such as *Keq*. Usually the product of linguistic field research, data content is any element that is essentially unanalyzed. (Though textual data in practice will contain implicit analysis, e.g., in the form of phonemic segmentation.) Lines 2 and 3 contain elements of data analysis, known more generally as **annotation**, or anything value-added that is not data content. Examples 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. We distinguish **terms** used to label elements of annotation from the abstract linguistic elements themselves, or those entities posited to exist by the linguist. That is, what are actually given in line 2 of the example 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 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, two terms can often have the same intended meaning, especially across different analyses. Finally, the example itself is given as **interlinear glossed text** (IGT), a common data structure or organizational device for grouping together data content and analysis for some presentational or computational purpose. We argue that most, if not all, data have these three components: the data content itself, the components of data analysis, and components of the data structure. This bears directly on being able to treat all kinds of data in a unified framework.

3. Components of the GOLD Community

GOLDComm consists of two sets of resources: those composed of linguistic data, and those composed of generalizations over that data. Data resources are the core of the model and represent its empirical part, one in which data are represented both in their raw form and in semi- or fully annotated form. The latter set of resources capture the collective knowledge of the field, that which is ultimately grounded in data.

3.1. Data-centric Components

3.1.1. *Best-practice Resources*

Based on Bird and Simons (2003b), 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 GOLDComm 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 approach 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 than are ones with highly structured models, 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.

In terms of particular XML structures, we follow already established practices (e.g., the E-MELD School of Best Practices [emeld.org/school/]) and encourage the use of DTDs or Schemas that are focused more on description and less on presentation. The main reason for preferring descriptive over presentation-centric XML is that adequate presentation can always be derived from well described content. Descriptive content is in a sense more fundamental than its presentational form, since the same content can be rendered in a number of different ways. Consider, for example, that whereas the entries in traditional print dictionaries are ordered according to alphabetic or similar orthographically-based criteria, those same dictionaries could be presented according to rhyming patterns, root morpheme, or even frequency. The point is that presentation follows from description and not vice versa, and that data should ultimately be maintained in a descriptive format. The rendering of descriptive data into a presentation-centric format is best considered as a separate application that can be built around GOLDComm. In fact, we argue that rendering is one of the key applications that will ensure the Community's success. If data are renderable in a multitude of different presentation formats, then many different groups can access the data in ways that make sense for them. This is particularly important when considering the dilemma often encountered in linguistic fieldwork, namely, how to balance the needs of the scientific community with the needs of the speaker community. Linguistic research demands a presentation organized according to particular theories, while the speaker community could be better served with a presentation organized, for example, to benefit language learners.

3.1.2. *Termsets*

One of the primary goals of GOLDComm 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, beyond manual effort. Instead, the mapping will be 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 some 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 resource.

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: A **termset** is a tuple $\langle T, C\rangle$, where:

1. T is a set of markup elements and C is a set of classes or instances from GOLD;
2. For each $t \in T$, there is zero or more $c \in C$ such that t denotes c;
3. 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* could represent 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 within the same termset to represent the same element in GOLD. Note that in a given termset, it should not be a stringent requirement that every term be mapped to an element of the ontology. Without the mapping, we would still consider such a resource to be a well-formed termset, however, the data described by the undefined term 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.

A termset is intended to be used as input to an automated processor for migrating the data to an interoperable format. Mapping to an ontology, other than simply providing semantic grounding, facilitates such applications as ontology-driven 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 encourages 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.

3.1.3. *Descriptive Profiles*

Termsets merely define the meanings of a terms within a particular context, e.g., within the scope of a certain document, for a particular language, for a particular research tradition, etc. 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 linguistic description is meant 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, or **grammar fragment**. A grammar fragment is defined as a formalization of some portion of a language's grammatical system.

Definition 2: A **grammar fragment** is a tuple $\langle C, L_{DS} \rangle$, where:

1. C is a set of linguistic concepts;
2. L_{DS} is a set of formal data structures;
3. 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 the co-occurrence constraints on morphosyntactic features. 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. Next, we introduce a 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 constructing profiles of lexicons, we propose that a descriptive profile include a termset and one or more grammar fragments.

Definition 3: A **descriptive profile** is a tuple $\langle T_s, G \rangle$, where:

1. T_s is a termset;
2. G is a grammar fragment;
3. The annotation in each $g \in G$ is expressed using T_s.

First, there is a termset indicating what the markup elements mean. The termset is followed by a grammar fragment that provides structure and constraints on some portion of the grammar. Profiles are useful within GOLDComm, because they facilitate the derivation of *new* knowledge from best-practice resources, for example, via the application of an implication given the presence of a particular feature.

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 un-structured formats, such as HTML or plain text documents, or in proprietary formats, including PDF and various word-processing formats which cannot be read in the absence of special software (e.g., Microsoft Word). Many such formats, especially the proprietary ones, are notoriously difficult to process automatically (Bird and Simons, 2003b). Generally speaking, the linguistic data structures used in these formats are presentation oriented, designed to accommodate the needs of presenting the data in a human readable format.

A major advantage of GOLDComm 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 GOLDComm. But *in lieu* of costly software to directly migrate legacy to best-practice resources, a kind of shortcut can be achieved by migrating only the most important, descriptively relevant aspects. This migration requires constructing a descriptive profile for the resource, which contains a terminology mapping and any relevant grammar fragments. A proof of concept of such a migration has been carried out in Simons et al. (2004), and the Online Database of Interlinear Text (ODIN) (Lewis, 2006) [www.csufresno.edu/odin/] implements this on a larger scale, with over 35,000 instances of IGT harvested from scholarly documents posted to the Web. A graphical summary of the various data-centric components is given in part A of 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 about that data. 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, these components capture the general, 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 techniques of these fields can be used to model specific scientific domains, e.g., the biomedical domain (Rosse et al., 2005). One of the key problems that GOLDComm addresses is the control and separation of various knowledge components. As will be discussed

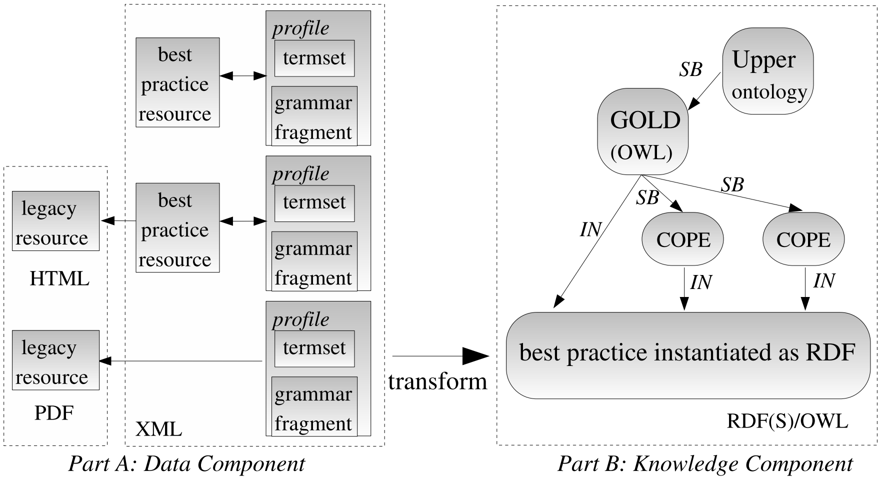

Figure 1. Graphical view of the GOLD Community

in the following section, the design provides a means of separating (1) general linguistic knowledge from (2) knowledge of particular languages and from (3) knowledge that pertains only to specific sub-communities of practice. Furthermore, the design allows for the relating of linguistic knowledge in GOLDComm to an upper ontology. In short, GOLDComm is 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 GOLDComm is GOLD, as introduced in Section 1. 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 to be 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 potentially 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”, all of which can be encoded in an expert or knowledge-based systems. The modeling choices in GOLD are described elsewhere (e.g., Farrar, In press); therefore here, we only mention a few key aspects of its implementation relevant for GOLDComm. 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 GOLDComm.

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 imperfective aspect or that English and Greek both have a past tense constitute linguistic knowledge; but, this kind of knowledge can be differentiated from the general knowledge referred to 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 for capturing 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., constructs of set theory or primitive relations such as subclass or instance), 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).

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 identifiable 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 across the data components to achieve a high degree of interoperability. But linguists not only employ different surface terminologies, they actually conceptualize the discipline 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: A **COPE** is a tuple $\langle C, I \rangle$, where:

1. C is a set of classes and I is a set of individuals in the COPE;
2. $\forall x.C(x) \exists y.C(x) \rightarrow C_g(y)$, such that C_g is a class in GOLD;
3. $\forall x.instance(x, C) \exists y.instance(x, y) \rightarrow C_g(y)$;
4. If K_i and K_j are COPEs, then $K_i \cup K_j$ 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 K_i is being created for a given language family and there already exists a language-neutral COPE K_j for the grammatical category of aspect. Then, some of K_i's members could be included in K_j. A key reason for having COPEs is to allow for contradictory knowledge to exist side by side in GOLDComm. Thus, while a COPE by definition cannot include knowledge that contradicts GOLD, two or more COPEs may contradict each other.

Thus, we envision several types of COPEs. First, consider a COPE for work from a focused descriptive tradition. For example, a description of languages such as Swahili and related languages, a focused and an extensive knowledge pertaining to Bantu noun classes is required. Such a COPE would facilitate the definition of a Bantu proto-noun class system and could be shared across the Bantu community. Second, the conceptualizations of some subdisciplines of linguistics could be constructed and maintained relatively separately, thus aiding in the management of the collective knowledge of the field. Consider for instance, the subdiscipline 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 conceptualizations encountered in linguistics, for example, Minimalism and Systemic Functional Grammar. Whereas some of the basic conceptualization is shared, e.g., the existence of LinguisticFeatures, 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 in theory-specific COPEs as mixing them with other traditions would likely introduce contradiction into any knowledge base.

Above all, the use of COPEs furthers the modular design of GOLDComm. A graphical summary of the various knowledge components is given in part B of Figure 1. From the figure, the main relation of subsumption (SB), which holds 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.

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 terminology. 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, but 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. This obviates 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 applications using GOLDComm, e.g., query engines. That is, if a user wants to exclude certain language-specific knowledge from their queries (if the analysis is questionable, or irrelevant), then by having this knowledge separated into various COPEs, the exclusion can be done in the query component by simply de-selecting a particular data source.

Finally, to address who will make up the GOLD Community, the data component discussed in this section will be built from the bottom up by anyone interested in data interoperability and access through best-practice markup. And GOLDComm provides a concrete framework to make this happen. Data is valuable precisely because they can be compared to other data within particular theoretical frameworks. Once researchers see the benefit of such a framework, namely applications built around the knowledge component, we believe they will be persuaded to include their data. The analogy is the Web itself. People include information on the Web, because it is a popular medium and because they are rewarded by having their content widely dispersed. More concretely, we look towards the success of the E-MELD and DoBeS, two projects that have been successful at promoting the notion of best practice and in building communities surrounding common tools and encoding schemes specifically for data from endangered languages. Finally, the only requirements for being included in the GOLDComm is to use a best-practice encoding and to link markup terminology to the GOLD ontology.

4. Comparison with Related Work

The GOLDComm effort is related to a number of other projects. First is the Open Language Archives Community (OLAC) (Simons and Bird, 2003) that is designed to enable the discovery of linguistic resources. The registered resources are discoverable because data providers are required to provide rich metadata that describe the contents. The **controlled vocabularies** used in the metadata can be seen as the precursors to our own use of a much richer resource, an ontology, to achieve more precise semantics, not just for resource metadata (Bird and Simons, 2003a), but in our case also for the data instances themselves. The GOLD Community and OLAC are indeed similar in many respects. For instance, GOLDComm relies on an extensive ontology of linguistic concepts to achieve interoperation, while OLAC relies on the OLAC Metadata Set which serves a similar purpose. The difference is that GOLDComm emphasizes the retrievability and interoperation of individual data instances, not whole documents, databases, or tools as OLAC does. Thus OLAC search is designed to be coarse. As an example of this, if one performs a search using the string "accusative", the OLAC Search Engine returns exactly one hit: "Walmatjari: Nominative-ergative or nominative-accusative?" While there are surely more data that refer to the accusative, the OLAC metadata does not indicate this, and the hit originates only from the title of the resource. It is likely in the future, however, that OLAC's search facility will be greatly improved by including even more metadata.

Also related to our work is the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 2002) that has created a standard format for data exchange within the humanities including linguistics. The TEI has adopted XML and established a core set of markup elements as well as elements to fit various specific domains within the humanities. These are defined within a set of established DTDs that are immediately accessible. One of the principles of the TEI that is meant to facilitate data interchange is the separation of content from presentation through the use of stylesheets. But we argue that our work takes the idea of interchange a step further. The main difference between the TEI and our work is that while the TEI seeks to achieve interchange through structural standardization and the use of common markup elements, GOLDComm provides a mechanism for migrating data with comparable structure onto a semantically interoperable resource. We argue that standardized XML structures are not enough to achieve interchange to the degree that we envision. Even so-called descriptive markup elements, e.g., <acc> indicating accusative case, are only useful when there is some way to interpret them in the context of other competing markup schemes. Also, TEI's scheme for terminological databases is similar to GOLDComm's termset. The obvious difference is TEI's inclusion of prose definitions instead of mappings to concepts in an ontology. Thus, it is not necessary to nest entries in the GOLDComm termset in the TEI sense, since the ontology provides the relevant hierarchical relations to structure terms that are related in some way.

Another difference concerns data types. Whereas the TEI, in many cases at least, seeks to provide very general data types that are meant to be broadly applicable across the humanities, GOLDComm requires **fundamental data types** of traditional linguistics (e.g., IGT, lexicons, paradigms). It has been our experience that simpler models for the fundamental data types are preferable to the ordinary working linguist. Indeed the TEI includes DTDs for print dictionaries, feature structures (and systems), syntactic trees, and language corpora, and these recommendations are certainly adequate for their intended purposes. The feature structure specification is particularly usable within GOLDComm off-the-shelf. However, there are no DTDs for paradigms or IGT. This is not necessarily a critique of the TEI, but may explain its lack of widespread adoption by linguists. Also related to the issue of data type is the lack of a strong unifying concept for linguistics data in the TEI. Data types in the TEI are treated quite separately, even though many dictionaries, often those of more “exotic” languages, contain IGT or other complexities such as higher degrees of recursion as observed in agglutinative languages (Weber, 2002). Such cases strain TEI's <gramGrp>, used for grammatical information.

Finally, a more general difference is that the TEI is fundamentally text-based as its name implies, though there are certainly aspects of the TEI that provide deeper analytic mechanisms: for linguistics, the feature structure construct is a prime example. The purpose and spirit of the TEI seems to fit the markup paradigm of transforming legacy material to digitally interchangeable resources. GOLDComm, on the other hand, seeks to provide a framework for the post-textual, in fact, Web-centered world, a world in which digitization and the use of computationally oriented data structures are assumed from the outset. A case in point is TEI's focus on print dictionaries as compared to more modern, machine readable lexicons. Using print dictionary DTDs goes against the GOLDComm's focus on content vs. presentation. Indicative of TEI's lack of focus on content is that there is no possibility to link to an ontology, something that is becoming more important in the design of on-line lexical resources (Calzolari et al., 2001).

Also relevant for our work here is ISLE (International Standards for Language Engineering) (Calzolari et al., 2003), a project that focuses on developing best practices and standards for HLT. ISLE builds upon and is largely coordinated with the EAGLES (Expert Advisory Group for Language Engineering Standards) initiative. ISLE had three working groups focusing respectively on computational lexicons, natural interaction/multimodality, and evaluation. Of particular relevance here is the working group on lexicons which produced recommendations for the Multilingual ISLE Lexical Entry (MILE), a kind of meta-entry, or common representational layer allowing interchange between specific projects. Thus, "MILE can be used to provide a means of communication and cooperation between those communities engaged in content-oriented description and access to services (semantic web, agent-based services, ontologies, content providers, . . .) and those engaged in overcoming the language barrier . . . " (Calzolari et al., 2003, p. 13). MILE Lexical Objects consist of three types of more basic objects: Lexical Classes, Lexical Data Categories and Lexical Operations. Thus, MILE emphasizes the importance of data typing and the necessity of having well-defined data structures. Furthermore, the MILE work shows how a Lexical Data Category Registry can be implemented in RDF (Ide et al., 2003). The work on MILE is relevant to GOLDComm, precisely because it includes recommendations on how to include reference to an ontology: "word-senses are encoded as Semantic Units or SemU. Each SemU is assigned a semantic type from the Ontology . . . " (Calzolari et al., 2002, p. 20). One ontology the ISLE recommends is the SIMPLE ontology (Lenci et al., 2000).

The ISLE project demonstrates that enormous complexities are involved in the standardization of even one linguistic data type, namely computational lexicons. The MILE work in particular covers much that is common to that of GOLDComm, including linguistic modeling in general, the representation of data on the Semantic Web through the use of RDF, and the use of ontologies. ISLE's use of ontologies, however, is slightly different than what is envisioned with GOLDComm. In MILE, the ontology provides a semantic reference point for the meanings of lexical entries, whereas all data elements in GOLDComm (including elements of phonology, morphosyntax, etc.) are linked to concepts in the ontology. However, it could be argued that MILE itself is a kind of ontology – in fact it is a schema – for much of what is covered in the GOLD ontology itself. At this time MILE is not formalized at the same level of detail as GOLD.

Thirdly, our work has many similarities with the Linguistic Annotation Framework (LAF) that is under development by a working group associated with ISO TC37 SC4 (Ide and Romary, 2004). The primary aim of the LAF is to provide a common format and abstract data model to which disparate data can be mapped regardless of representation. This is referred to as the **dump format**, described as "isomorphic to the data model and intended primarily for machine rather than human use" (Ide and Romary, 2004). The dump format assumes a strict separation between structure and content. Structurally, the dump format consists of a feature structure graph which contains a number of **data categories** as content. A commonly used notion in ISO 12620, a data category is the result of a specification of a given data **field**, that is, an element of annotation appearing in individual language resources (Romary, 2003, p. 5). Examples cited in Romary (2003) are gender and part of speech. Data categories have particular **data elements** associated with them, such as masculine and feminine for the gender category. The semantics of each of the data categories and elements is defined according to a prose definition in a **data category registry** (Romary, 2003). A data category registry can contain standardized, or user-defined, data categories. Crucially, however, the data categories are registered and are, therefore, easily sharable among the community. "The DCR [data category registry] is intended to provide a set of formally defined reference categories" (Ide and Romary, 2004), which ensures that categories are at least well defined, presumably by experts. The DCR is currently being used, for example, to construct lexical entries in the LEXUS tool (Kemps-Snijders et al., 2006).

Without a doubt, the overall goals of the LAF are similar to those of GOLDComm. However, there are two key differences. First, the LAF relies on a semi-structured, standardized set of data categories, the data category registry. Our approach, on the other hand, takes advantage of knowledge expressed in a logic with a formal semantics, organized according to a formal ontology. It should be noted that a data category registry as described by Ide and Romary (2004) can be considered a *light-weight* ontology of sorts, as it includes well defined data categories and (implicit) relations over them. It would be perfectly in step with our general proposal to output DCR elements and build a COPE that is meant to capture the canonical data categories used in linguistic annotation. In this way, the knowledge captured by the DCR could be cast as knowledge of a deeper sort, resulting in more precise search and deeper automated reasoning capabilities. GOLD is furthermore related to an upper ontology, thus opening up the possibility of positioning large amounts of linguistic data within a broader scientific context. Second, the LAF data model is focused on feature structures. Our proposal on the other hand incorporates feature structures as one of several data structuring alternatives. In fact, by using the ontology, it is possible to focus solely on data content abstracting away from theory-specific data structures. But even with these differences, we see such standardization efforts as complementary to GOLDComm. Since the registries encourage consistency in the form of quasi-standard uses of terminology by experts in particular subfields, starting with such a controlled vocabulary would allow users to more easily migrate data to a knowledge-rich format just as with GOLDComm.

5. Putting the Model to Use

Turning to the question of how GOLDComm can be put to use. The first and perhaps most important application that GOLDComm will facilitate is **ontology-driven search** over massive amounts of semantically disparate data. There are essentially two types of ontology-driven search envisioned within GOLDComm: **concept search** and **intelligent search**. 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 current 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 likely to return documents concerning Pacific Standard Time! On the other hand, a more sophisticated concept search for Subject would return data that are marked for all of the following: *Subject*, *SUBJ*, *NOM*, and *ERG* (ErgativeCase). Another example of concept search, demonstrated in Simons et al. (2004), is: “List language data for all languages where one word encodes both past tense and second person." 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 preterit, a form of the past tense; the morpheme *monuhmon* marks *2Conj*, a form of second person; and both morphemes are in the same word. Of course real concept search, for linguistics at least, is not yet possible, though GOLDComm is intended to set the stage for its eventual realization.

A less sophisticated form of concept search has been implemented in ODIN (Lewis, 2006). Using ODIN, users can select from a list of GOLD concepts, and find IGT that contain instances of morphemes encoding these concepts. For example, a search for SingularNumber will return all IGT examples that contain morphemes glossed *SG*, as well as those glossed *SING*, *1SG*, *2SG*, etc. Although not as sophisticated as the proof-of-concept search implemented in Simons et al. (2004), ODIN is publically available online and boasts a search facility across data for over 700 languages.

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 with a case appropriate to object position. In nominative/accusative languages like Yaqui, such a noun would be marked for accusative case. Thus, the search actually performed is "List all instances of nouns marked for accusative case in Yaqui that are arguments of the verb".

Our proposal, then, leverages the ever increasing body of digitized resources to move linguistic inquiry into a new era where automated analysis is not only possible, but in fact *de rigueur* in any empirical research program. It is, we argue, only through the use of codified knowledge resources such as GOLD that this vision can be achieved.

Acknowledgements

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 E-MELD project [emeld.org] (NSF ITR-0094934). 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. Finally, we gratefully acknowledge the NSF-funded Data-Driven Linguistic Ontology Development project (BCS-0411348) which supported the authors during the writing of this manuscript.

References

Aristar, A.: 2003, 'FIELD'. Technical report, presented at the 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 G. F. Simons: 2003a, 'Extending Dublin Core metadata to support the description and discovery of language resources'. *Computers and the Humanities* **37**, 375–388. http://www.arxiv.org/abs/cs.CL/0308022.

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

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

Calzolari, N., F. Bertagna, A. Lenci, and M. Monachini: 2002, 'Standards and Best Practice for Multilingual Computational Lexicons & MILE (the Multilingual ISLE Lexical Entry)'. ISLE Deliverable D2.2-D3.2, ISLE Computational Lexicons Working Group. http://www.ilc.cnr.it/EAGLES96/isle/clwg_doc/ISLE_D2.2-D3.2.zip (2006-07-09).

Calzolari, N., R. Grishman, and M. Palmer: 2001, 'Survey of major approaches towards Bilingual/Multilingual Lexicons'. ISLE Deliverable D2.1-D3.1, ISLE Computational Lexicons Working Group, Pisa.

Calzolari, N., J. McNaught, M. Palmer, and A. Zampolli: 2003, 'ISLE D14.2-Final Report'. ISLE Deliverable D14.2, ISLE. http://www.ilc.cnr.it/EAGLES96/isle/ISLE_D14.2.zip (2006-07-09).

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. www.u.arizona.edu/∼farrar/papers/Far-fc.pdf.

Farrar, S. and D. T. Langendoen: 2003, 'A linguistic ontology for the Semantic Web'. *GLOT International* **7**(3), 97–100. http://www.u.arizona.edu/∼farrar/papers/FarLang03b.pdf.

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

Ide, N., A. Lenci, and N. Calzolari: 2003, 'RDF Instantiation of ISLE/MILE Lexical Entries'. In: *Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right*. Sapporo, pp. 30–37. http://www.cs.vassar.edu/∼ide/papers/ACL2003-ws-ISLE.pdf (2006-07-09).

Ide, N. and L. Romary: 2004, 'International standard for a linguistic annotation framework'. *Journal of Natural Language Engineering* **10**(3-4).

Kemps-Snijders, M., M.-J. Nederhof, and P. Wittenburg: 2006, 'LEXUS, a web-based tool for manipulating lexical resources'. In: *LREC 2006: Fifth International Conference on Language Resources and Evaluation*. pp. 1862–1865.

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. http://www.u.arizona.edu/∼farrar/papers/LangFarLew02.pdf.

Lenci, A., F. Busa, N. Ruimy, E. G. M. Monachini, N. Calzolari, and A. Zampolli: 2000, 'Linguistic Specifications'. SIMPLE Deliverable D2.1, ILC and University of Pisa, Pisa. http://www.ub.es/gilcub/SIMPLE/reports/simple/SIMPLE_FGuidelines.rtf.zip (2006-07-09).

Lewis, W. D.: 2006, 'ODIN: A Model for Adapting and Enriching Legacy Infrastructure'. In: *Proceedings of the e-Humanities Workshop held in cooperation with e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing*. Amsterdam.

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. http://home.earthlink.net/ adampease/professional/FOIS.pdf.

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.

Rosse, C., A. Kumar, J. L. M. Jr, D. L. Cook, L. T. Detwilern, and B. Smith: 2005, 'A Strategy for Improving and Integrating Biomedical Ontologies'. In: *Proceedings of AMIA Symposium 2005*. Washington, DC, pp. 639–643.

Simons, G. and S. Bird: 2003, 'The open language archives community: An infrastructure for distributed archiving of language resources'. *Literary and Linguistic Computing* **18**, 117–128. http://www.arxiv.org/abs/cs.CL/0306040 (2006-May-17).

Simons, G. F., W. D. Lewis, S. O. Farrar, D. T. Langendoen, B. Fitzsimons, and H. Gonzalez: 2004, 'The semantics of markup: Mapping legacy markup schemas to a common semantics'. In: *Proceedings of the 4th Workshop on NLP and XML (NLPXML-2004): Held in cooperation with ACL-04*. Barcelona, Spain, pp. 25–32. http://www.u.arizona.edu/∼farrar/papers/Sim-etal04b.pdf.

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

Weber, D. J.: 2002, 'Reflections on the Huallaga Quechua dictionary: derived forms as subentries'. In: *On-line Proceedings of the 2002 E-MELD Workshop on Digitizing Lexical Information*. http://saussure.linguistlist.org/cfdocs/emeld/workshop/2002/presentations/weber/emeld.pdf (2006-07-07).