The GOLD Community of Practice

An Infrastructure for Linguistic Data on the Web[1]

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. A number of services around the model are proposed including knowledge acquisition and search facilities. Finally, as an example of the model's utility, an instantiation of a community of practice centered around interlinear glossed text is described.

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) which was introduced in Farrar, Lewis & Langendoen (2002) and described in more detail in Farrar & Langendoen (2003) and Farrar (forthcoming). 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

[1]Presented at the Summer 2005 *E-MELD Workshop on Morphosyntactic Annotation and Terminology: Linguistic Ontologies and Data Categories for Linguistic Resources*, July 1–3, Radcliffe Institute, Harvard University. For an expanded version of this paper, see http://www.u.arizona.edu/ farrar/papers/FarLew-prep.pdf

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, Hendler & Lassila 2001), the Community provides the means whereby linguists can use diverse terminology, yet arrive at a consensus through what the terms mean and, thus, achieve true data interoperability.

The paper is structured as follows: 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 services built around the model. In Section 5 we describe our experiences with a particular community of practice formed around the use of a particular linguistic data structure commonly referred to as 'interlinear glossed text'. Finally, in Section 6 we provide a discussion and summary.

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 linguistics descriptive literature and illustrates some key features of linguistic data.

(1) Yama+ ajurnu jarti- ki +karn pikapikka- ka ngu- nngara.
leave+ 3PL.NS other- DAT+ now children- DAT lie- OPT.FUT
'Some should be left for the other children.' (Simpson 1998, p. 727)

Line 1 from (1) contains actual data content, in other words, linguistic expressions such as *Yama*, while lines 2 and 3 contain elements of description and analysis, such as *3PL* or the free translation. (Line 1 also contains analysis, i.e., phonemic/segmentation analysis.) The example itself is given as **interlinear glossed text** (IGT), a very common linguistic data structure (discussed further in Section 5). 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, namely, the GOLD Community of Practice. We now expound upon these three aspects of linguistic data.

The data content itself is usually the product of linguistic field research and is found in a variety of media formats including textual, audio, and video data content. Data content is any element that is essentially unanalyzed. That is, we contrast elements of data content with elements of data analysis. We do not assume that data content is 'primary'. That is, data content will have been transcribed or recorded by a linguist, and therefore cannot be 100% true to the original linguistics events produced by the native speaker. Given current best technology, a high quality video recording of a conversation can be considered as close to the actual event as possible. For many purposes, an audio recording is a sufficiently detailed rendition of the event. Any textual transcription, however, incorporates an element of analysis by the linguist. A phonetic transcription of a speech event can be used to capture much of the detail needed for further analysis, but it can never be fully true to the original. Even so, for the purposes of this model, we consider such video, audio and phonetic transcriptions as data content. The major problem with data content is that it often exists outside of any particular context, though limited context information often accompanies data content in the form of **metadata** (IMDI 2003).

Anything that is not data content belongs to the realm of data analysis. Examples of data analysis include a phonemic transcription of a speech event (captured by line 1 of Example 1), 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 & Liberman (2000) have called 'annotation'. Elements of data analysis are often theory-specific, as the same data content can be analyzed in multiple ways. However, as we will argue later, there are quasi-standard constructs that linguists generally agree upon. Other than specific grammatical categories, morphemes, phonemes, and semantic representations are well-known constructs used for analysis. Other examples of linguistic analysis include the positing of constructs such as 'inflectional unit', 'syntactic constituent', or 'synonym'. Analyses may also be particular to the description of an individual language or to a certain grammatical tradition. Consider the tradition in the description of the Bantu languages of labeling the noun classes with numerals, e.g., 'Class-15' in Swahili, or the traditional use of 'masculine' and 'feminine' to refer to grammatical gender in Romance languages.

Referring to (1), we distinguish terms used to label elements of analysis from the elements themselves (those entities posited to exist either in the mind of speakers or shared across a speech community). That is, what is 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,[2] but this 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, Farrar & Lewis 2002). On the one hand the same term can be used ambiguously. For example, the term *NOM* could be used to label either 'nominative case' or 'nominalizer' (see Section 5). On the other hand, different terms can often have the same intended semantics, especially across different analyses. One possibility for dealing with diverse terminology in linguistics is to develop scientific standards (e.g., ISO/TC 37); however, 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 [3] 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]The meaning of markup elements, viz. analysis, should not be confused with the linguistic meaning of natural language expressions.

[3]See http://www.ethnologue.com/.

Finally, a data structure is an organizational device for grouping together data content and analysis for some display, theoretical, or computational purpose. That is, a data structure is an information container which provides a useful way to package information. The quintessential example of a linguistic data structure is IGT, as in (1) with a very long tradition going back at least to the structuralists. IGT is a very common device used to structure data content with gloss information and phonological, morphological, and/or syntactic analysis. Another common data structure is the dictionary entry, used to group an arbitrary amount of information about a word of some language. Some other examples of these are given in Table 2.1. Generally speaking, such data structures are **display oriented**, designed to accommodate the needs of displaying the data in a human readable format. These display-oriented devices have a long tradition in the field of linguistics and play a key role in the readable presentation of data in scholarly publications. The main characteristic of display-oriented data structures is human readability. It is common of such a data structure, for example, to group information into tables or lines of text.

display-oriented	**non-display-oriented**
phonetic chart	phonological rule
phonology tableaux	feature structure
dictionary entry	phrase structure rule
comparative word list	lexical entry
morphological paradigm	verb frame
syntactic tree	logical representation

Table 1: A list of a common data structures

At the other end of the continuum are the data structures required for linguistic theories and computational systems. Linguistic theories and computational systems often require much more formal structures than the display oriented ones mentioned above. Some familiar data structures already in wide use in the theoretical and computational communities are given in the second column of Table 2.1. The main feature of such non-display-oriented data structures is that they are either specific to a particular linguistic formalism or designed for direct implementation on a computer. As a case in point, compare a display-oriented 'dictionary entry' to the computationally-oriented 'lexical entry'. (Here, we use the terms *dictionary* and *lexical* merely to bring out the difference between display and computational approaches.) A dictionary entry can certainly be a highly structured object complete with sub-entries and cross-references. However, the primary intent is for human readability, and thus may contain an *ad hoc* amount of information. A lexical entry for some grammatical formalism on the other hand is more constrained. It is constrained by the particular theory to include only a fixed amount of information, and the information is in general strongly type-contrained.

Characteristics of data structures in general include having recursive properties, meaning that a particular structure can be embedded with another, as in the case of a lexical entry containing a feature structure or a verb frame containing a lexical entry. Furthermore, data structures can be directly linked to other mathematical formalisms such as set theory, lattice theory, grammar writing algebras, and systems of symbolic logic. For the purposes of our work, the link to symbolic logic is particularly important since the ontology is also specified in a symbolic logic.

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 & Burnard 2002) and also the work of Bird & Simons (2003). At a minimum these principles require the consistent use of Unicode (The-Unicode-Consortium 2000) and XML (Bray, Paoli, Sperberg-McQueen, Maler & Yergeau 2004) to encode and markup data content. XML contrasts, for example, with less structured formats including HTML and text documents and with proprietary formats, e.g., Microsoft Word (discussed further in Section 3.1.4).

In terms of the elements of data analysis, E-MELD and Bird & Simons (2003) recommend mapping all terminology to a semantic resource, an ontology, that defines them. Note that this recommendation is meant to avoid the direct promulgation of standard terminologies. It was realized early on in the E-MELD project for example, that consensus regarding terminology was very difficult even within specialized communities. It was therefore proposed (e.g., in Farrar et al. (2002)) to map all terms in a domain-specific ontology for linguistics. We will say no more about this now, since it is this topic, in fact, that is the focus of the current work.

Finally, there exists already several proposed best-practice standards for various types of linguistic data structures. A proposal for lexicons, for example, has been given in Bell & Bird (2000). In a similar vein, proposals for IGT and linguistic paradigms have been given in Hughes, Bird & Bow (2003) and Penton, Bow, Bird & Hughes (2004) respectively. These works provide broad survey information for each data structure in question and then provide a best-practice recommendation focused on XML. Finally, Lee (2003) offers a proposal for the best-practice markup of typed feature structures based on the previous TEI-oriented work of Langendoen & Simons (1995).

2.3 The Vision of the Semantic Web

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 exists no hard and fast solution to creating a Semantic Web. However, it is our claim that, at least for science, the Semantic Web is achievable when approached from the bottom up. That is, it is the responsibility of individual 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 achieve the Vision.

Rather than echo the Vision statement here, we opt to give a short introduction into the technologies of the Semantic Web that we will use to develop the GOLD Community of Practice. The introduction here assumes a knowledge of two of the most basic technologies related to the Semantic Web: XML and Unicode, which were described in Section 2.2 as key to best-practice encoding and markup of linguistic data. XML is lauded in part because it is extensible, that is, versatile enough to be used in a myriad of different applications and structured just enough to be used a data exchange format. While XML is versatile and structured, its ability to express explicit relationships within or across documents is particularly weak. This is due to the fact that XML makes very little commitment to what elements actually mean.

A partial answer to the problem of explicitly expressing meaning is the Resource Description Framework (RDF) (Lassila & Swick 1999). In RDF a 'resource is anything that can be identified

on the Web, such as a webpage, or an individual XML element or value – in short, anything with a Uniform Resource Identifier (URI). RDF provides the mechanism to make statements about Web resources; it is an information model meant to convey that something represents something else. RDF provides the syntax for basic entity-relationship graphs with explicitly named edges, or relations. RDF is a structuring format (a syntax) for markup and is itself serializable in XML. A semantic extension of RDF called RDF Schema (Brickley & Guha 2004) adds the ability to form classes of resources and to define relations according to what kinds of classes can be in the domain and range.

The development of XML and RDF(S) is only the first step in assigning meaning to Web content. The second step includes the formation of semantic resources which will act as repositories of knowledge about particular domains, linguistics for example. The primary type of semantic resources for the Semantic Web are ontologies, which enable the current Web to be transformed into a highly structured knowledge system. A Web ontology is intended to be much more than just a controlled vocabulary, that is, more than just a pre-defined set of terminology. Ontologies on the Semantic Web will require explicit representation of ontological commitment and axiomatization. According to (Masolo, Borgo, Gangemi, Guarino, Oltramari & Schneider 2002), the ontologies envisioned for the Semantic Web fall into two categories: the so-called 'lightweight ontologies (often just taxonomies) that provide "semantic access to a specific resource", where the meanings of terms are already known; and 'foundational ontologies that are to be used to help Web agents "negotiate meaning" or to establish a consensus between human users and agents. The current W3C recommendation for representing ontologies is the Web Ontology Language, or 'OWL' (McGuinness & van Harmelen n.d.). OWL adds expressibility to RDF(S) by providing ways to define complex classes, e.g., based on how they relate to other classes or according to what kinds of relations they participate in. There are various dialects of OWL, though our work will only concern OWL-DL, which can be classed as a type of 'description logic' (Baader, Calvanese, McGuinness, Nardi & Patel-Schneider 2003).

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 4.1 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 is 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 display-centric, proprietary resources which are not in XML. The following section describes each of these types of data resources while attempting to emphasize

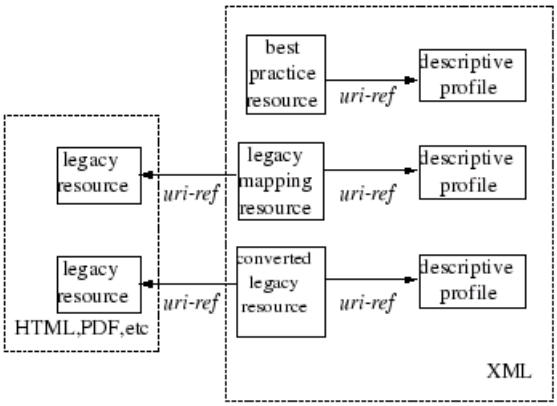

Figure 1: Data-Centric Components of the GOLD Community

the need for best-practice.

3.1.1 Best-practice Resources

Based on Bird & Simons (2003) and the discussion in Section 2.2, we adopt the general notion 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 concern 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. The model could accommodate more structured data resource formats, e.g., RDF(S) or OWL, but these formats are more appropriate for implementing the knowledge components (to be discussed in Section3.2). But for a general data format, we argue that a pure 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 work with and a broad variety of software is available for editing XML documents. We argue, then, 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 & Simons (2003).)

In terms of particular XML structures, we encourage the use of DTD 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 display can always be derived from well described content. Descriptive content is in a sense more stable than display, since the same content can be rendered in a number of different ways. Consider, for example, the ordering of entries in an electronic dictionary, where traditional print models focus on alphabetic or similar orthographically-based ordering. But a dictionary can be presented in a number of different ways: entries can be grouped according to rhyming patterns, root morpheme, or even frequency. The point is that display follows from description and not vice versa. The rendering of data into a display-centric format is best considered as a service that can be built around the GOLD Community of Practice. In fact, we argue that rendering is one of the key services 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 encounter 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 the knowledge-based components. Even with well designed best-practice resources in place, the mapping process would be a daunting task and in most cases precludes manual effort. The mapping should necessarily be (semi-)automated. But any hope of automation requires something beyond best-practice. One of the primary reasons for this 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 them. Some terms 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 terms mean. Therefore, we suggest the use of **termsets** to accompany 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 terms 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, $\{c_1, c_2, \ldots, c_n\}$, then interpret the set as the union $c_1 \cup c_2 \cup c_n$.

Although it is possible for a term, even within a limited community, to represent more than one concept (for example, *NOM* representing either 'nominative case' or 'nominalizer'), we require that only terms with a conjunctive meaning, e.g., *3SG* or *1PL*, be used in this manner. Note that this does preclude the use of identical terms in two or more disjoint data resources, where the two do not share the same termset. Finally, it is allowable for multiple terms to represent the same element in GOLD. Appendix A shows a snippet of an XML implementation of a termset that conforms to the definition. Note that the term *CL* in Appendix A is not defined according to the ontology, thus not conforming to the requirement that each term is 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. 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 in Section 4.1. Mapping to an ontology, other than simply providing semantic grounding, facilitates such services as "smart" or concept-based search. For example, a query for the concept SINGULAR would return data described with *SG*, as well as *1SG*, *2SG* and *SING*, etc. Furthermore, 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 & Romary 2003) and (Ide, Romary & la Clergerie 2003).

One aim of ISO TC37 SC4 is to develop 'data category registries' (Romary 2003). The advantage 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 be useful for the markup of lesser studied languages.

3.1.3 Descriptive Profiles

Termsets are, in a sense, snap-shots of the grammar of a language 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 a 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 descriptive linguistics is intended to capture. Therefore, 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; that is, a grammar fragment is a data structure as defined according to best-practice principles. In the GOLD Community of Practice, this can include any kind of useful grammatical information, e.g., the possible morphophonemic combinations of a language or co-occurrence constraints on morphosyntactic features (see Section 2.2). A grammar fragment is just that, a *fragment*, because the knowledge of a language's structure, function, etc. will almost always be incomplete.

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 formal data structure;
- Each $c \in C$ is contained in some $l_{DS} \in L_{DS}$.

So, with termsets and grammar fragment in place, we introduce the next type of data centric resource, a **descriptive profile**. Inspired by work on the FIELD tool (Aristar 2003), we propose that a descriptive profile include a termset and one or more grammar fragments.

Definition 3 (Descriptive Profile). A descriptive 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.

Appendix B shows a descriptive profile with the minimal requirements. 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 Appendix B, 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 cases values are exhaustive.

3.1.4 Legacy resources

A major advantage of the GOLD Community of Practice is that it accommodates not only best-practice resources, but also Web resources as they currently exist, namely those that do not conform to best-practice. Most computer-based language documentation uses off-the-shelf, general purpose software, such as Word or Excel (Bird & Simons 2003), formats that are notoriously difficult to process automatically. In the quest for a common platform for disseminating data and analyses, linguists have increasingly relied on PDF as a data format for information interchange. Such a move is well-motivated and on the surface would appear to solve potential data-interchange problems. Unfortunately PDF, although quite portable, is strictly display-centric, and it can be argued is even less automation friendly than the data formats from which it is generated. We refer to any Web resource of this type – ones that do not conform to best practice – as a **legacy resource**. Most linguistic data on the Web at the moment reside in legacy resources. Legacy resources are either in unstructured formats, such as HTML and text documents, or proprietary formats, including PDF and various wordprocessing formats which cannot be read in the absence of special software. Bird & Simons (2003) stress that data encoded in proprietary formats is particularly at risk, since its long-term survival relies on the existence and continued support of software in which it is encoded. It should be noted that formats such as HTML and PDF, although widely supported and thus less likely to become obsolete, are primarily used for human consumption and suffer from the problem of being display oriented and not readily translatable.

The abundance of Web data encoded in such formats suggests that a mechanism for migration or access would be of immense importance to the field. We discuss the migration of legacy formats in Section 4.1.1, but note here that *in lieu* of costly software to directly migrate legacy to best-practice resources, a kind of short-cut can be achieved. Instead of migrating the entire contents of the legacy resource (data and analysis), we advocate migrating only the most important, descriptively relevant aspects. This migration would involve constructing a profile for the resource, which would encode a terminology mapping and any relevant grammar fragments. Since the descriptive information contained within a profile would hold crucial facts that facilitate query, such resources could be located and queried over, even if the resources themselves are not in best practice format or even accessible. At such time that resource migration becomes desirable or feasible, the profile document could help in the migration process by providing the necessary terminology mapping, thus supplying a means for modularizing the migration process (i.e., structural vs. semantic).

3.2 Knowledge-centric components

With a description of the basic data-centric components out of the way, we now turn to a description of the components concerned with knowledge. The purpose of the knowledge-centric components is to represent explicitly the knowledge that is captured implicitly by the data. On the one hand, the knowledge-centric components capture the general, canonical knowledge of the field. On the other, they represent 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 formal and applied ontology. One of the key problems that the GOLD Community addresses is the control and separation of various knowledge components, shown graphically in Figure 2.

As be discussed in the following section, the design provides a means to separate general linguistic knowledge from the knowledge of particular languages and from knowledge that pertains only to specific subcommunities of practice. Furthermore, the design allows for relating the linguistic knowledge of the GOLD Community to an upper ontology. In short, we provide an implementation of the vision of the Semantic Web.

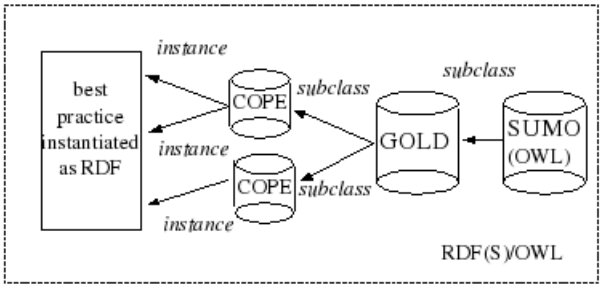

Figure 2: Knowledge-centric components of the GOLD Community

3.2.1 The General Ontology for Linguistic Description

The most central component of the GOLD Community of Practice is the **General Ontology for Linguistic Description** (GOLD). Thus far, in the description of data-centric components, we have focused on that individual linguistic descriptions, that which is common to a particular theory or a 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 potentially forms the basis of any theoretical framework; but, in partiuclar, GOLD captures the fundamentals of descriptive linguistics. Examples of such knowledge are given below:

- A verb is a part of speech.
- A verb can assign case.
- Gender can be semantically grounded.
- Linguistic expressions realize morphemes.

The list above shows knowledge of a generic sort, that which is typically represented in ontologies in expert and knowledge-based systems. The modeling choices in GOLD are described elsewhere (see Farrar (forthcoming)), therefore here, we only mention a some key aspects of its implementation relevant for the GOLD Community of Practice. We note here that, whereas GOLD could be used to represent linguistic universals, e.g., in the sense of Greenberg (1966), we do not include them. Instead the derivation of implied universals could be given as a potential service on top of the GOLD Community of Practice.

In practice it is difficult to separate general linguistic knowledge from that pertaining 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 that listed above, as it only pertains to specific language. This issue relates to the problem of differentiating between theoretically-specific knowledge and that which pertains to the entire field. We do not claim that GOLD is completely theory independent, but claim that its categories are at least applicable to a diverse set of linguistic theories. In the next section we provide a solution to the problem of how to keep such sorts of knowledge separate.

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. To capture explicitly the relationship of sub-community knowledge to GOLD, we include a type of 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 categories depending on the specific requirements of the sub-community, and that extends GOLD in ways particular to the sub-community. We envision several parameters around which communities of practice within linguistics can be constructed:

- theoretical perspective
- language or language family
- data type (e.g., a community centered around interlinear glossed text)
- terminology

We can now give a more precise definition of a community of practice within our Model: it is the consistent use of a particular COPE. With COPEs, communities have the ability to maintain the knowledge central to their community in discrete, manageable packets. This provides at least two benefits. From an implementation point of view, 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 profile documents, the exclusion can be done in the query component by simply deselecting a particular data source. And second, from a knowledge engineering perspective, COPEs can be mined to add missing knowledge to GOLD. For instance, if GOLD is lacking a particular concept, then it will show up in numerous different COPEs. The knowledge 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.

3.2.3 The Upper Ontology

A final concern of the GOLD Community of Practice is how to link the knowledge amassed for linguistics to the rest of the sciences, in particular, ones which may also have established a Semantic Web presence. The answer is to link GOLD with an **upper ontology**. An upper ontology provides the means of linking linguistic constructs to a larger conceptual framework. In fact, an upper ontology is necessary for the creation of GOLD in the first place, since it provides the basic tools for constructing the an ontology for linguistics: 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. To this end we are currently considering two upper ontologies: the Suggested Upper Merged Ontology (SUMO) (Pease & Niles 2002, Niles & Pease 2001) and the Descriptive Ontology for Language and Cognitive Engineering (DOLCE) (Masolo, Borgo, Gangemi, Guarino & Oltramari 2003). We will not describe these ontologies here, though interested readers are referred to Farrar & Bateman (2004) for a discussion and evaluation of various upper ontologies.

4 Services Based on the GOLD Community of Practice

Thus far, the discussion has focused on a description of the GOLD Community from a relatively static point of view. However, we envision that the Community will provide a dynamic 'workplace' for linguistic data. In order to achieve this, we propose various services built on and around the infrastructure. According to their primary functions, the services can be divided into two groups: those used for knowledge acquisition and those used for knowledge use. Both types of services pose significant challenges for computer science and knowledge engineering. We discuss each type in the following section.

4.1 The Path from Data to Knowledge

To begin with services used for knowledge acquisition, recall from Section 3.1.4 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 access to them. Certain kinds of semi-structured data are common in linguistic discourse and are often encapsulated in documents encoded in proprietary file formats. These include many of the data display-oriented data structures in Table 2.1. Automated extraction or migration of these data types from proprietary file formats provides some potential for leveraging existing resources into a richer data format, such as XML. If the XML format conforms to best-practice recommendations, then a migration to knowledge-centric components is readily achieveable. And, if done on a large enough scale, the full migration could help to ensure the acceptance of the model presented here.

4.1.1 Legacy to Best-Practice

Because of the number of proprietary file formats used by linguists, and the even larger number of formats linguists use to encode their data, it is not possible to discuss legacy conversion in specific detail for all the formats that are used. We do, however, propose the following general design guidelines for migrating legacy data:

1. Construct profiles or termsets that define the semantics of terminology, tags, and symbols used, along with any other descriptive information that may be useful for migrating, processing, or accessing the data.
2. Use existing tools to migrate to an intermediary form and ensure minimal data loss.
3. Build tools to map the intermediary form, which includes element alignment and font transforms, to XML.

We postpone giving an example of such a migration till Section 5 where we discuss the migration of IGT that is encapsulated in PDF documents. And for more discussion of automated conversion of particular types of semi-structured data formats from legacy file formats, see Simons, Fitzsimons, Langendoen, Lewis, Farrar, Lanham, Basham & Gonzalez (2004) and Lewis (2003).[4]

However, since the idea is relatively new, the Community has to rely also on the large number of legacy data resources already available on the Web, ie. those not in best practice. It is the intention of the Community to promote services that transform legacy resources to best-practice.

[4]An implementation using migrated IGT can be seen at http://www.csufresno.edu/odin. The Online Database of Interlinear Text (ODIN) was funded as part of the Data Driven Linguistic Ontology Grant.

In the actual implementation, legacy resources will be mapped to a set of descriptive profiles. This is shown in Figure 1.

We recognize that it may not always be feasible, or in some cases possible, to migrate legacy resources to a best practice format. As noted in Section 3.1.4, creating a descriptive profile for a resource can act as a stop-gap for legacy data that cannot yet be migrated. The data can still be located, and depending on the richness of the profile, queried or even reasoned over. A profile then acts as one step in the process of migration from data to knowledge.

4.1.2 Best-practice to knowledge

As already noted, the GOLD Community of Practice is founded on data encoded as best-practice resources. But given that the data may come from a variety of disparate sources, about different languages, described from different theoretical perspectives, it is necessary to map these data onto a common semantic resource – GOLD. But going from data to knowledge is not a simple transformation and requires an advanced Web service. For example, various terminologies used in the best practice resources first need to be rendered transparent and compatible with one another by mapping them onto a set of descriptive profiles. These resources in turn allow for the transition from semi-structured data (XML) to highly-structured knowledge (RDF(S)/OWL).

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. 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 SIL, then, 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 consisting 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, viz. the OWL ontology. Furthermore, the metaschema formally interprets the original markup structure by declaring what the dominance and linking relations in the XML document structure represent. The metaschema process is implemented using XSL transformations. We have demonstrated that the migration process can be successfully implemented in a scalable, systematic fashion. For more details on the SIL and the notion of the metaschema, see Simons, Fitzsimons, Langendoen, Lewis, Farrar, Lanham, Basham & Gonzalez (2004) and Simons, Lewis, Farrar, Langendoen, Fitzsimons & Gonzalez (2004).

4.2 Putting the Knowledge to Use

We finally come to the question of how the GOLD Community of Practice can be put to use. The first and perhaps most important service that the GOLD Community will provide is smart search over massive amounts of disparate data. In what follows we describe our experience with building linguistic databases and implementing smart search as services built around the GOLD Community of Practice.

4.2.1 Building Linguistic Databases

As discussed in Section 4.1.2, creating best-practice resources themselves is not sufficient to instantiate the Model: the ontology is a critical part of the puzzle, ensuring a common semantics for interpretation and interoperation. Equally important are the services that make the enrichment

of resources possible, as are those that provide the capacity to transform existing and resulting structures into a knowledge-rich representation that best facilitates interoperation, including, but not limited to, smart search. Thus, building terminology sets and profiles and migrating data to a knowledge-rich representation are essential services in the model, and essential for building linguistic databases of the size and diversity envisioned.

A small scale implementation of a diverse linguistic database is demonstrated in Simons, Lewis, Farrar, Langendoen, Fitzsimons & Gonzalez (2004), where we construct a linguistic database from a variety of resources, including:

- Four sets of interlinear data mined from scholarly papers encoded in PDF or Word (Archi, Hausa, Hopi, Passamaquoddy)
- One lexicon with interlinearized forms (Hopi)
- One set of interlinear data mined from an HTML format (Warrungu)
- One annotated treebank corpus (Korean)
- One Shoebox file (Mutsun)
- One XML encoded data file (Monguor)

Each resource existed in or was converted to XML, and paired with each one was a termset and profile. Using SIL, all data and analyses were converted to a common RDF form, and the resulting RDF output was then read into a Sesame database (Jeen Broekstra & van Harmelen 2002) for the purposes of query. Queries of various sorts were posed against the database, many comparing data and analyses across the various languages and data types included in the database.

This project demonstrated the potential for building larger databases from heterogenous data types and for building services to migrate data. SIL is a precursor to such services, as were the customized tools used to convert legacy data to best-practice XML (e.g., those that converted the Shoebox file and interlinear data formats). Not demonstrated in this paper was a service for constructing termsets and profiles; these were created by hand. Although termsets are not particularly complicated to build by hand – termsets can be thought of as lists of term and concept pairs – creating profiles is a more rigurous process. The FIELD tool (Aristar 2003), which outputs complex language profiles mapped to GOLD, provided the template for the profiles created in Simons, Lewis, Farrar, Langendoen, Fitzsimons & Gonzalez (2004), and can be considered a precursor to services constructed for such purposes. The FIELD profile builder acts as a rudimentary expert system, directing users through a series of questions, guided by GOLD, and constructs a profile document containing descriptive grammar fragments informed by the answers to these questions.

4.2.2 Smart Search

The primary advantage of enriched linguistic databases is the ability to perform smart searches across data and analyses. There are essentially two types of smart searches envisioned within the GOLD Community, those that we call *concept* searches, and those that we call *intelligent* searches. The former uses the ontology to inform searches whereby users specify the concept they are looking for. 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. String searches can only return instances that match the input, character for character, and are uninformed as to the meaning of the string specified. They all too frequently generate too much output, in which the string expresses an unexpected ambiguity, or too little output, in which the string does not

capture instances that are semantically similar but named differently. 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).[5] An example of concept search used in Simons, Lewis, Farrar, Langendoen, Fitzsimons & Gonzalez (2004) is: "List language data for all languages where one word encodes both PastTense and SecondPerson." The query returned an instance of data from the Passamaquoddy IGT data set, the only instance that satisfied the condition (shown in 2). 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.

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

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, and 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 Experience with Interlinear Glossed Text

As a test case for the GOLD Community, we have explored the creation of a sub-community centered around the use of 'interlinear glossed text' (a data type that was briefly illustrated in Section 2.1). In doing so, we have shown that it is possible to accomplish a semi-automated conversion of legacy resources to a more interoperable, best-practice format. We have also demonstrated that current legacy resources implicitly represent a community of practice, one which is centered around a common terminology used in IGT. We have also demonstrated the utility of building termsets (defined in Section 3.1.1) and of augmenting termsets with language-specific or resource specific knowledge to create descriptive profiles.

5.1 Overview of Interlinear Glossed Text

Interlinear Glossed Text (IGT) is a common method for encoding and presenting linguistic data, and is often used in scholarly papers to present snippets of language data relevant to a particular analysis. The most common form is the three-line format, a sample of which is shown in (3) below.[6]. The first line gives data for the language in question, either phonetically encoded or in the language's native orthography. The line is broken down into words and morphemes, where words are usually delimited by spaces, and morphemes by dashes ("-"), although other characters

[5]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.

[6]For an overview of the variety of interlinear formats used in linguistics, refer to Hughes et al. (2003)

can be used to delimit morphemes, such as "+" and "=" (the latter two are most often used for delimiting clitics). The second line contains a morpheme-by-morpheme or as word-by-word gloss for the data in the first line, or a mixture of the two. The second-line delimiters are generally the same as those in the first, and morphemes and words usually align between the two lines. Where a given word or morpheme can be glossed by more than one term – for instance, *kataab* is glossed as *book*, *3s* ('third person' and 'singular number'), *f* ('feminine gender'), and *NOM* ('nominative case') – periods (".") or colons (":"), and sometimes dashes or spaces, separate the additional glosses. Finally, the third line contains a free-translation of the first line.

(3) fawad- ne sumbal- n kataab dittii
Fawad- ERG Sumbal- DAT book.3s.f.NOM give-PERF.f
'Fawad gave the book back to Sumbal.'(Akhtar 1997)

Glosses in the second line take two forms: those representing grammatical information (usually formal or semantic features), which we label with the term *gram* (in the spirit of Bybee & Dahl (1989), and described in more detail in Lewis (2003)), and those that contain unconstrained translations of words and morphemes, which Lewis (2003) labels by the generic term *gloss*. Grams are often put in upper case to differentiate them from glosses.

Although IGT is a very rich data structure, and presents a significant amount of data in a very limited amount of space (which probably explains the continuing popularity of IGT as form of linguistic data presentation), what makes IGT difficult for automated processing is its bias towards presentation, with much implied in its structure. For instance, the alignment between the first and second lines is done through delimiters (as described), and the only way to automatically align a morpheme with its gloss is by counting these delimiters. Further, since stems are not overtly identified in the format, determining what is a prefix or what is a suffix is often left to the reader, and generally cannot be discerned automatically (without employing more complicated methods of tagging, parsing, and aligning text). Misalignments, such as cases where more dashes occur in second line versus the first, can also complicate automated processing. Finally, terminology is a problem since, although there are *de facto* standards for what grams used in IGT mean (see the next section), these standards are often violated, and there can be multiple meanings for any given term, and multiple terms for any given meaning. All this being said, however, the structural and terminological conventions shared by most IGT can be leveraged for building a database of interlinear text within the GOLD Community of Practice, built from extant IGT found on the Web.

5.2 Building a Database of IGT Found on the Web

It is becoming common for linguists to post scholarly papers to the Web, with documents appearing on linguists' own Web sites, in special online publications (such as Snippets), or even within online versions of linguistic publications (such as Language and Oceanic Linguistics through Project Muse). Because of this trend, a signficant volume of IGT is also finding its way to the Web in the form of examples included in such documents. At the time of this writing, the authors have discovered and mined over 40,000 instances of IGT from hundreds of scholary documents, to a total of about 400 languages' worth of data, and are leveraging these data to build an IGT database which will serve as a test case for the GOLD Community Model. Such a data store can serve as a good test of the Model not only because of the sheer volume of data, but also because of terminological conventions observed for the data type, and equally, because of the terminological differences where they occur. The former can be used to develop an IGT COPE, which would define the standards used across the data type, and the latter for the construction of resource, community, and language specific termsets and profiles. Because we have automated the process of mining and extracting IGT from PDF documents, and have developed techniques to migrate

interlinear data to a best practice data format (following the XML model for Interlinear text described in Hughes et al. (2003)), we have been able to build a foundation of data upon which tools and services can be developed. The database we are constructing we have called ODIN (the Online Database of INterlinear text).[7] In the succeeding section, we describe the COPEs, termsets and profiles that can be built around this database.

5.3 IGT COPE, Termsets and Profiles

Although defining a community as being 'users of interlinear glossed text' may seem too coarse a grouping to be useful, linguists as a group, irrespective of theoretical leaning or language family studied, are surprisingly consistent in the terms they use in IGT (Lewis 2003). For instance, the term *NOM* almost always refers to 'nominative case'; *ACC* almost always refers to 'accusative case'; *3SG* almost always refers to 'third-person singular', etc. Table 2 shows a list of the most common terms used in IGT and their most common meanings.[8]

Term	# of Tokens	# of Files	Likely GOLD Referent
ACC	2157	292	AccusativeCase
NOM	2698	279	NominativeCase
PAST	1255	208	SimplePastTense
DAT	1140	206	DativeCase
PL	887	206	Plural
GEN	816	180	GenitiveCase
1SG	1076	168	FirstPerson, Singular
3SG	1142	168	ThirdPerson, Singular
NEG	1145	146	Negation
SG	1434	142	Singular
LOC	423	115	LocativeCase
PERF	1498	112	PerfectiveAspect
FUT	455	111	AbsoluteFutureTense
PRES	616	109	AbsolutePresentTense
ERG	687	109	ErgativeCase
3PL	420	101	ThirdPerson, Plural
CL	850	98	Clitic or Classifier
TOP	523	96	Topic
2SG	299	95	SecondPerson, Singular
INF	353	77	Infinitive
PASS	210	72	PassiveVoice
REFL	245	71	Reflexive
1PL	179	66	FirstPerson, Plural
ABS	512	66	AbsolutiveCase
CAUS	268	63	CausativeVoice

Table 2: Top 25 terms in IGT (1,200 linguistic documents)

Even though local inconsistencies will occur—and these should be defined locally (i.e., in termsets and profiles)—many of terms used by all IGT resources could be defined within one centrally housed IGT COPE (Community of Practice Extension, as defined in section Section 3.2.2). The IGT COPE would allow interoperation across resources that are defined as IGT, in many ways

[7] http://www.csufresno.edu/ODIN

[8] It may be worthwhile to see Bickel, Comrie & Haspelmath (2004) for a larger list of IGT terms and their respective meanings. Although the list is meant to prescribe use, many of the terms listed are common in IGT, and some coincide with the terms shown in Table 2.

simplifying the migration and use of such resources. Further, one central IGT COPE affords the capacity to redefine ontological referents should the ontology change or other ontologies become available or usable, and would additionally allow the definition of terms specific to IGT or that may have no ontological referents. The COPE then reduces duplication of effort and affects the overall utility of IGT resources.

In building ODIN, we concentrated on constructing the IGT COPE first, ahead of defining local termsets and profiles. The COPE has been a helpful device for setting default concept values for terms, and has aided in IGT data mining, since newly discovered instances of known terms can be automatically mapped to the 'global' concept values defined in the COPE. In those cases where local terminological uses apply, the default values can be overridden within a termset. The metric we have used to determine whether a term should be defined in the COPE has been based on frequency: Only those cases where one meaning predominates significantly over the other is an entry made in the COPE. In cases where no such predominant meaning can be discerned, no entry is made. For instance, we discovered that the term *CL* is used as often for the notion 'clitic' as it is for the notion 'classifier'. In this case, no entry for *CL* appears in the COPE.

Although a similiar ambiguity exists for the term *NOM*, the significant frequency difference between meanings dictated its inclusion. As described earlier, the predominant meaning for *NOM* is 'nominative case'. Although such a default will work in most cases, there are rare instances where *NOM* is used to mean 'nominalizer'. This use can be seen in (4):

(4) tɕile- ŋuəɳi pə - s pies ŋuə.
1pl- TOP buy- NOM meat COP
'What we need is meat.' (LaPolla 2003)

Rather than excluding *NOM* from the IGT COPE—as was done with the term *CL*—which would affect the interoperation over the majority of IGT where *NOM* actually refers to nominative case, the nominalizer usage was defined locally within a termset for the resource or resources that use the alternate meaning. An example is shown in Appendix C. Note in particular that the term *1pl* is not defined in the termset since the term is already defined in the IGT COPE, and the LaPolla's usage is consistent with the definition provided there. The other terms, namely *TOP* for 'topic' and *COP* for 'copula', are also defined locally since they are not defined in the COPE. Interoperation over these terms is not limited since both terms are still defined by mapping to the ontology.

We have also developed descriptive profiles for resources within which IGT has been discovered. As described in Section 3.1.3, descriptive profiles extend termsets by providing *grammatical sketches* in addition to terminology-to-ontology mappings. Descriptive profiles that accompany resources can give exhaustive lists of possible grammatical features for specific parts of the grammar. Possible inventories include complete lists of case values, tense values, aspect values, etc. We have found that profile generation can be semi-automated given a set of terms extracted from IGT for a particular language or set of resources. For example, in instances where the case inventory extracted for IGT for a language is complete, we have created a grammar fragment in a profile that lists the possible cases for that language. Likewise for tense, aspect, voice, modality, etc. The process cannot be fully automated since it requires user intervention to determine if a particular inventory is in fact complete. Depending on the amount of data, and the use the data was put to, an inventory of a cases, for instance, could have accidental gaps. The language is not necessarily missing a particular case, it just has not been found in the source descriptions.

Supervision is also necessary to tease apart conflicting analyses for a particular language. For instance, perhaps Linguistic A claims that Language X has NominativeCase and it is marked on nouns. In turn, perhaps Linguistic B claims that Language X has ErgativeCase and it is marked on both nouns and pronouns. A profile across the resources from both Linguist A and B might

actually compose analyses that may not be compatible with one another. Discovering this fact is difficult, would require some degree of supervision, and would necessarily require the creation of two profiles for the language, one for each linguist. Once done, however, the *described varieties* represented by these profiles could be compared and contrasted.

It is important to point out that termsets and profiles can be created for resources that are not themselves in best practice. We have found that full conversion to best practice format for IGT resources was not necessary to extract the relevant information. In some cases, migration was not feasible (for instance, where alignment across tiers was not consistent), yet the knowledge contained within these instances was still retrievable, and could be compiled into a profile. The profiles so created are just as useful as profiles created for best practice resources, in that they can still be searched and the source resources located. Interoperation ends for these resources, however, at the level of data.

An example of a minimal descriptive profile, with a complete list of attested case feature values derived from IGT for the language Georgian, is shown in Appendix B.

6 Summary and Discussion

We have presented a model for a community of practice centered around linguistic data on the Web and an ontology for linguistic description, namely GOLD. The model was designed with the nature of linguistic data itself in mind. It was inspired from recent efforts to establish best-practice encoding and markup schemes, especially that suggested by Bird & 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 which can be 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 using profiles is a way to create a tangible artifact used for the creation of specific communities of practice, centered around a consensus about what terms mean.

7 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 grant 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 (NSF-0411348) which supported the authors during the writing of this manuscript.

A Sample Termset

```
<profile ID="Georgian">
   <semanticResource>
     <prefix ID="gold-morph"/>
     <resource ID="http://www.linguistics-ontology.org/ns/gold/0.2/gold-morph.owl">
     <prefix ID="gold-pos"/>
     <resource ID="http://www.linguistics-ontology.org/ns/gold/0.2/gold-pos.owl">
   </semanticResource>
    <terms>
       <term ID="ERG">
          <concept>gold-morph:ErgativeCase</concept>
          <comment>Ergative Case marks the pronominal subjects of transitive verbs.</comment>
       </term>
       <term ID="GEN">
          <concept>gold-morph:GenitiveCase</concept>
          <comment>GEN marks the possessive on nouns.</comment>
       </term>
      <term ID="N">
         <concept>gold-pos:Noun</concept>
         <comment>IMPF only occurs in the present tense.</comment>
      </term>
      <term ID="CL">
         <comment>CL marks nominal classifiers. Referent unknown.</comment>
      </term>
</termset>
```

B Sample Descriptive Profile

```
<profile ID="Georgian">
   <semanticResource>
     <prefix ID="gold-morph"/>
     <resource ID="http://www.linguistics-ontology.org/ns/gold/0.2/gold-morph.owl">
     <prefix ID="gold-pos"/>
     <resource ID="http://www.linguistics-ontology.org/ns/gold/0.2/gold-pos.owl">
   </semanticResource>
   <terms>
     ...
   </terms>
   <gram>
      <feature>gold-morph:Case
         <value>gold-morph:NominativeCase</value>
         <value>gold-morph:AccusativeCase</value>
         <value>gold-morph:DativeCase</value>
         <value>gold-morph:ErgativeCase</value>
         <value>gold-morph:GenitiveCase</value>
      </feature>
   </gram>
```

C Termset from LaPolla

```
<profile ID="LaPolla">
  <ontologyNamespace prefix="gold">http://emeld.org/gold</ontologyNamespace>
  <terms>
    <term>NOM
      <concept>gold:morpheme</concept>
      <value>gold:Nominalizer</value>
    </term>
    <term>TOP
      <concept>gold:morpheme</concept>
      <value>gold:Topic</value>
    </term>
    <term>COP
      <concept>gold:POS</concept>
      <value>gold:Copula</value>
    </term>
</termset>
```

References

Akhtar, R. N. (1997), 'Affix -s(uu) constructions in punjabi', *Essex Graduate Student Papers in Language and Linguistics* **1**.

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

Baader, F., Calvanese, D., McGuinness, D., Nardi, D. & Patel-Schneider, P., eds (2003), *The Description Logic Handbook*, Cambridge University Press.

Bell, J. & Bird, S. (2000), 'A preliminary study of the structure of lexicon entries', Presented at the Workshop on Web-Based Language Documentation and Description.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001), 'The Semantic Web', *Scientific American* .

Bickel, B., Comrie, B. & Haspelmath, M. (2004), The leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses, Technical report, Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig. Revised version.

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

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

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E. & Yergeau, F. (2004), Extensible markup language (xml) 1.0 (third edition), Technical report, World Wide Web Consortium (W3C).

Brickley, D. & Guha, R. V. (2004), RDF vocabulary description language 1.0: RDF Schema, Recommendation. Available at http://www.w3.org/TR/rdf-schema/.

Bruening, B. (2001), Syntax at the Edge: Cross-Clausal Phenomena and the Syntax of Passamaquoddy, PhD thesis, MIT.

Bybee, J. L. & Dahl, O. (1989), 'The creation of tense and aspect systems in the languages of the world', *Studies in Language* **13**(1), 51–103.

Farrar, S. (forthcoming), Using 'Ontolinguistics' for language description, *in* A. Schalley & D. Zaefferer, eds, 'Ontolinguistics: How ontological status shapes the linguistic coding of concepts', Mouton de Gruyter, Berlin.
www.u.arizona.edu/∼farrar/papers/Far-fc.pdf

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

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

Farrar, S., Lewis, W. D. & Langendoen, D. T. (2002), An ontology for linguistic annotation, *in* 'Semantic Web Meets Language Resources: Papers from the AAAI Workshop, Technical Report WS-02-16', AAAI Press, Menlo Park, CA, pp. 11–19. This paper references EMELD and had more exposure than any other publication.
www.u.arizona.edu/∼farrar/papers/FarLewLang02b.pdf

Greenberg, J. (1966), *Language Universals*, Mouton, The Hague.

Hughes, B., Bird, S. & Bow, C. (2003), Interlinear text facilities, *in* 'E-MELD 2003', Michigan State University. [http://emeld.org/workshop/2003/baden-demo.html]. See also [http://www.cs.mu.oz.au/research/lt/emeld/interlinear].

Ide, N. & Romary, L. (2003), Outline of the international standard Linguistic Annotation Framework, *in* 'Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right'.

Ide, N., Romary, L. & la Clergerie, E. D. (2003), International standard for a Linguistic Annotation Framework, *in* 'Proceedings of NAACL'03 Workshop of Software Engineering and Architecture of Language Technology Systems'.

IMDI (2003), IMDI (isle metadata initiative) part 1 metadata elements for session descriptions, Technical Report v. 3.0.4, Max-Plank Institute for Psycholinguistics.

Jeen Broekstra, A. K. & van Harmelen, F. (2002), Sesame: A generic architecture for storing and querying RDF and RDF schema, *in* I. Horrocks & J. Hendler, eds, 'Proceedings of the First International Semantic Web Conference', Springer-Verlag, Berlin, pp. 54–68.

Langendoen, D. T., Farrar, S. & Lewis, W. D. (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.
www.u.arizona.edu/∼farrar/papers/LangFarLew02.pdf

Langendoen, D. T. & Simons, G. F. (1995), 'A rationale for the TEI recommendations for feature-structure markup', *Computers and the Humanities* **29**, 191 – 209.

LaPolla, R. J. (2003), Dulong, *in* R. J. LaPolla & G. Thurgood, eds, 'The Sino-Tibetan Language', Routledge, New York. http://victoria.linguistlist.org/ lapolla/rda/.

Lassila, O. & Swick, R. R. (1999), Resource description framework (rdf) model and syntax specification, Recommendation. Available at http://www.w3.org/TR/REC-rdf-syntax/.

Lee, K. (2003), (draft) language resource management—feature structures—Part I: Feature structure representation, Technical Report ISO TC 37/SC 4 N033 Rev. 1 2003-07-25, ISO. Available at http://www.tei-c.org/Activities/FS/iso-n033-1.pdf.

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., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. (2003), Ontologies library (final), WonderWeb Deliverable D18, ISTC-CNR, Padova, Italy.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. & Schneider, L. (2002), The WonderWeb library of foundational ontologies: preliminary report, WonderWeb Deliverable D17, ISTC-CNR, Padova, Italy.

McGuinness, D. L. & van Harmelen, F. (n.d.), OWL web ontology language: Overview, organization = W3C, type = Recommendation, month = Feb, year = 2004, note = Available at http://www.w3.org/TR/owl-features/, Technical report.

Niles, I. & Pease, A. (2001), Toward a standard upper ontology, *in* C. Welty & B. Smith, eds, 'Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)', Association for Computing Machinery, Ogunquit, Maine.
http://home.earthlink.net/adampease/professional/FOIS.pdf

Pease, A. & Niles, I. (2002), 'IEEE Standard Upper Ontology: A progress report', *Knowledge Engineering Review: Special Issue on Ontologies and Agents* **17**.
http://home.earthlink.net/∼adampease/professional/KER.ps

Penton, D., Bow, C., Bird, S. & Hughes, B. (2004), Towards a general model for linguistic paradigms, *in* 'Presented at the EMELD'04 workshop on databases and best practice'.

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.
http://www.sil.org/∼simonsg/metaschema/sil.htm

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., Fitzsimons, B., Langendoen, D. T., Lewis, W. D., Farrar, S. O., Lanham, A., Basham, R. & Gonzalez, H. (2004), A model for interoperability: XML documents as an RDF database, *in* 'Proceedings of the EMELD Workshop on Databases', Detroit, MI.
www.u.arizona.edu/∼farrar/papers/Sim-etal04a.pdf

Simons, G. F., Lewis, W. D., Farrar, S. O., Langendoen, D. T., Fitzsimons, B. & Gonzalez, H. (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)', Barcelona, Spain, pp. 25 – 32. held in cooperation with ACL-04.
www.u.arizona.edu/∼farrar/papers/Sim-etal04b.pdf

Simpson, J. (1998), Warumungu (Australian–Pama-Nyungan), *in* A. Spenser & A. M. Zwicky, eds, 'The Handbook of Morphology', Blackwell, Oxford.

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

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

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