A linguistic ontology for the semantic web

By Scott Farrar and Terry Langendoen

1. Introduction

The **World Wide Web** has the potential to become a primary source for storing and accessing linguistic data, including data of the sort that are routinely collected by field linguists. Having large amounts of linguistic data on the Web will give linguists, indigenous communities, and language learners access to resources that have hitherto been difficult to obtain. For linguists, scientific data from the world's languages will be just as accessible as information in on-line newspapers. For indigenous communities, the Web will be a powerful instrument for maintaining language as a cultural resource. For students and educators, a new tool will be available for teaching and learning minority and endangered languages. For linguists in particular, having linguistic data on the Web means that data from different languages can be automatically searched and compared. Furthermore, the Web would provide ready computational resources for the development of machine translation and other multilingual tools.

However, simply posting massive amounts of linguistic data is not sufficient. Rather, the data and the various encoding schemes in which they are represented need an explicit semantics. What we envision is a data model that goes far beyond that of the Web as we know it today, but which is consistent with what has come to be known as the **Semantic Web** (Berners-Lee, Hendler, and Lassila, 2000). Once Web data have a universally recognized semantics, efficient tools can be created that will benefit all users. Moreover, **communities of practice**, groups with common goals and needs, can be expected to form around common tools, resources and encoding standards. We have taken the first step toward the creation of a linguistic community of practice by beginning work on a **General Ontology for Linguistic Description (GOLD)**, as part of a larger effort to create domain-specific ontologies connected to an **upper ontology** known as **SUMO** (see http://ontology.teknowledge.com).

Ontologies are central to the architecture of the Semantics Web, since they provide the basis for automated reasoning in a domain by declaring what entities exist and what kinds of relations hold between those entities. Such reasoning is enabled by markup languages, specifically the **eXtensible Markup Language (XML)** (see http://www.w3.org/XML/) and its extensions, including the **Web Ontology Language (OWL)** (see http://www.w3.org/2001/sw/), much in the way that HTML currently enables the display and linking of Web documents.

It is not necessary to adopt any particular linguistic theory in order to use GOLD or any other properly designed linguistic ontology. It only requires understanding what terms mean, and what can be inferred from their use in particular contexts, something that linguistics students are already routinely trained to do. In order to post, retrieve and analyze data on the Semantic Web, an ontology is needed that integrates different theories and terminologies. For example, if one data set contains the term 'Class 1' and another uses the combination of terms 'Human' and 'Singular', both referring to a noun class representing human individuals, then these two data sets can be equated by intelligent Web agents.

2. Creating a linguistic ontology

Creating a comprehensive ontology that can be useful to the future linguistic community of practice is a daunting task, even with the help of a broad upper model such as SUMO. We organize linguistically related concepts into four major domains: expressions, grammar, data constructs, and metaconcepts. In the next two subsections, we discuss our approach to **expressions** and **grammar**. Here we comment briefly on the other two domains.

Data constructs are constructs that are used by linguists to analyze language data, such as paradigms, lexicons and feature structures. We include data constructs because we need to relate disparately structured data resources using the ontology, for example, the relationship between the concept LEXEME in a lexicon to that of STEM in a paradigm. **Metaconcepts** are the most basic concepts of linguistic analysis, including language itself. Historically, linguists have had difficulty agreeing on what language is, undoubtedly because of the many ways in which language can be viewed. Every introductory linguistics text book has a definition of sorts and most linguists if pressed could come up with one. Without a working concept of language, an ontology cannot be used to describe and compare data from all of the world's languages. Since our primary aim is to compare **data**, we have defined a language as the

Scott Farrar and Terry Langendoen, Department of Linguistics, PO Box 210028, The University of Arizona, Tucson AZ 85721-0028, USA, farrar@u.arizona.edu, langendt@u.arizona.edu

set of data associated with a common grammatical pattern. Despite this inadequate and certainly controversial characterization of language, the definition allows us to move ahead. What is perhaps more crucial is that, in GOLD, language is related to the concept HUMAN in a precise way: Humans produce language, and the elements of language mean, express, embody, realize, code, represent, or symbolize something else (Payne, 1997, 6). Hence, language is representational which is expressed as a foundational axiom in GOLD.

We use two strategies in the construction of GOLD. One is bottom-up; we survey the properties of specific languages to ensure that the ontology has sufficient coverage. The other is top-down; we represent general, widely accepted linguistic universals, for example the relationship between closed- and open-class expressions and the fact that all languages have words that refer directly to instances of PHYSICAL-OBJECT. The goal is to represent the cumulative knowledge of well trained and broadly experienced professional linguists, who know about both individual languages and universals.

2.1 Expressions

By expressions, we refer to the physically accessible aspects of language, for example the printed words you are reading on this page and the sounds produced when you speak. SUMO already contains the basic segmental notions such as WORD, PHRASE and SENTENCE subsumed under LINGUISTICEXPRESSION. We have refined the concept by including under it WRITTENLINGUISTICEXPRESSION and SPOKENLINGUISTICEXPRESSION. Informally, WRITTENLINGUISTICEXPRESSION is the class of content bearing objects whose members make up the symbol set of a WRITTENLANGUAGE. Such linguistic symbols are non-abstract, non-mental things in the world such as strings of characters. The concept SPOKENLINGUISTIC EXPRESSION includes the sounds of a SPOKENLANGUAGE. Instead of being subsumed under OBJECT, these spoken forms are subsumed under PROCESS in SUMO. The resulting partial taxonomy is shown in (1). (Note: tabs represent the subclass relation.)

(1) Upper taxonomy for expressions in SUMO/GOLD

```
Entity
   Physical
      Object
         SelfConnectedObject
         ContentBearingObject
            Icon
            SymbolicString
               Character
               OrthographicString
            WrittenLinguisticExpression
      Process
         ContentBearingProcess
            SpokenLinguisticExpression
```

The relationship between a WRITTENLINGUISTIC EXPRESSION and a SPOKENLINGUISTICEXPRESSION is denoted by the EQUIVALENTCONTENTCLASS predicate with no commitment as to which concept is logically prior. Since our immediate goal is to apply GOLD to reasoning on the Semantic Web, we have chosen to focus on WRITTENLINGUISTICEXPRESSION. Future versions of GOLD will include phonetics and phonology and will necessarily include a complete characterization of SPOKENLINGUISTICEXPRESSION. A **partial** classification of written segments is provided in (2).

(2) Taxonomy of written segments in GOLD

```
WrittenLinguisticExpression
   WordPart
      SimpleWordPart
         Root
         Affix
            Prefix
            Infix
            Suffix
         Clitic
      Stem
   Word
      SimpleWord
      ComplexWord
         Compound
```

The segments in (2) are adequate to describe most written phenomena that we have encountered in the domain of morphosyntax. In particular, we do not want to proliferate categories of segments by creating new ones for each language. For example, the category SUFFIX does not subsume any other language-specific category such as WARUMUNGUSUFFIX or ENGLISH-SUFFIX. Our analysis of segments is based on the rigid mereological and positional properties of symbolic strings. That is, we classify the segments according to which ones are part of another and where those parts are located in relation to the whole. For instance, a segment at the terminus of a Word will be classified as a Suffix if that string is part of the inventory of morphemic segments in some language. Mereological relationships between segments are specified by axioms such as ''a stem has a root as one of its parts'', expressible in a suitable markup language such as OWL.

2.2 Grammar

By grammar, we refer to the abstract properties and relations of language, the domain that is of primary interest to linguists. To say we model grammar conceptually is tantamount to saying that we account for the very nature of language, a pursuit that has been going on for millennia and has yet to reach its goal. We are not making such grand claims nor are we proposing a new unified theory of grammar. Rather, we are simply proposing a framework in which individual grammatical phenomena and the various theories of grammar can be compared and related, and that can facilitate automated reasoning. Even in

these terms, modeling grammar is no small task. As a starting point, we have chosen to focus on the domain of morphosyntax.

We propose that anything expressed by a grammatical system be represented by the concept GRAMMATICALCATEGORY. Using a traditional metaphor, conceptual space is 'carved up' or 'discretized' by morphosyntactic and other grammatical categories (Payne, 1997, 51). Grammatical categories exist as internal properties of a language yet have a clear connection to reality or at least how reality is conceptualized. Essentially, we agree with Pollard & Sag (1987) regarding the relation between realism and conceptualism; they assume that grammatical categories represent **information**. We remain agnostic as to whether those categories arise from features of the external world (nominalism), features of the organism (conceptualism), abstract features (realism) or combinations thereof. For example, the morphosyntactic feature <Number singular> that describes a grammatical number property of some linguistic expression often implies that the physical object described by the linguistic expression has a cardinality of one, regardless of whether the users of the language conceptualize that object as such.

Grammatical categories are not physically accessible like linguistic expressions. The conceptual or abstract nature of the grammatical categories leads to some difficulty in both describing individual grammatical systems and comparing across them. As a concept, GRAMMATICALCATEGORY is on the same ontological level as the concepts COLORATTRIBUTE and SHAPEATTRIBUTE. An analogy using one of SUMO's preexisting categories is RED is to COLOR as PAST is to TENSE. However, whereas COLOR is an attribute of physical objects, TENSE is an attribute of a (construct of a) grammar, a conceptual or abstract object.

The morphosyntactic categories that we are interested in at this stage of the project are the ones commonly discussed in typology and morphology texts. A preliminary listing of some of these categories is given in (3).

(3) Some grammatical categories

Tense	Number	Voice	Valence
Mood	Person	...	Word Class
Aspect	Case		...
	Agreement		

In order to classify two or more of the categories in (3) under a single concept, it is necessary to develop axioms that can be used to form subcategories. While it may be intuitively clear that TENSE and ASPECT can be subsumed under one category, without an axiom to this effect, the grouping cannot be used for automated reasoning. The criteria we used for grouping grammatical categories are based, not on the morphosyntactic behavior of expressions which varies widely from language to language, but on what kinds of ontological entities they represent. So, for example, we are able to conflate TENSE and ASPECT in a common subcategory since they concern the category PROCESS as a whole as opposed to a category like ANIMATE that refers to some participant in a process. Essentially, we are agreeing with Greenberg (1963), Bybee (1985), Talmy (2000), and others who have shown that there is a relation between grammatical categories and certain non-linguistic categories, OBJECT, PROCESS, QUALITY, and SITUATION, that are part of the upper ontology. In short, our criteria for categorizing grammatical phenomena are based on what is most relevant to representing the semantics of linguistic expressions. Put another way the ontology is a tool for describing grammar in terms of meaning.

3. Related work

GOLD is the first ontology being designed specifically for linguistic description on the Semantic Web. Two types of projects, however, have provided inspiration for the development of GOLD. First are typology projects that bring together large bodies of language data, such as Autotyp (Bickel & Nichols, 2002), the Typological Database Project (Monachesi et al., 2001), and WALS (Dryer et al., forthcoming). We see the role of such typological databases as providing quality control for the information about grammatical structure that will become increasingly available over time. Web access would make these resources maximally beneficial to the linguistic community. But without a central resource such as GOLD, such databases will become increasingly opaque to outside comparison. Thus, GOLD can act as a kind of lingua franca for the linguistic data community provided that data providers are willing to map their data to GOLD or to some similar resource. What separates GOLD from other computer-based typology projects is that it is based on the principles of **knowledge engineering**. That is, domain knowledge is made maximally explicit in a knowledge representation language, enough, for example, to apply the power of an expert system to reason about language data and to apply the accumulated knowledge of a well-trained linguist.

Our second sources for inspiration are the lexical and ontological projects that are being developed for computational linguistics and natural language processing (NLP). While not a formal ontology, WordNet (Fellbaum, 1998) is a lexical resource that is rich enough to be considered alongside actual ontologies. WordNet contains an extensive taxonomic and mereological structure which could be regarded as a kind of 'proto-ontology'. Although Gangemi et al. (2002) demonstrate that a substantial transform of WordNet's upper categories is needed in order for it to be used directly as an ontology, every noun and verb in WordNet has now been mapped to categories in SUMO (see http://ontology.teknowledge.com). Other linguistically motivated ontologies have a substantial presence in computational linguistics and NLP. The most comprehensive such ontology to date is the Generalized Upper Model (GUM) (Bateman et al., 1994). GUM is designed at ''a level of abstraction midway between surface linguistic realizations and

'conceptual' or 'contextual' representations'' of natural language (Bateman, 2001). As such, GUM is intended to be used for NLP tasks, in particular language generation, and not for general reasoning, although it does have many categories in common with GOLD. Its categories and relations are linguistically motivated in that they include only those concepts necessary for processing language but partially conform to the organization of knowledge in general. For example, it includes the mereological concepts necessary to process a sentence like *gravel is an ingredient of concrete* and concepts such as NameOf necessary for interpreting sentences such as *the ship is called Knox*. Thus, GUM is an attempt at an intermediate level of abstraction bridging the gap between linguistic and non-linguistic knowledge. Most recently the WonderWeb Project described in Masolo et al. (2002) attempts to bring together a number of ontologies for natural language processing on the Semantic Web. One of these is the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), which focuses on the categories ''underlying natural language and human commonsense'' (Gangemi et al., 2002, 2). These resources, then, were developed primarily for NLP. GOLD, however, is an attempt to merge a rich knowledge of language and language data with the extra-linguistic knowledge already encoded in SUMO–all this for the primary purpose of reasoning about the wide variety of the world's language structures and not as a tool for processing English or Spanish texts.

We see GOLD then as a solution to the shortcomings of both typological database projects, namely the lack of interoperability with other projects, and projects designed specifically for NLP applications, which have traditionally not focused on linguistic data per se. The development of GOLD as a part of the Semantic Web is a logical next move that will both foster cooperation among linguists, indigenous communities, and language educators, and insure that crucial language data are not lost.

Acknowledgements

This work is supported in part by a National Science Foundation grant number 0094934, Electronic Metastructure for Endangered Language Data (EMELD), to the LINGUIST List at Eastern Michigan University and Wayne State University; see http://emeld.org/. The University of Arizona is a subcontractor on this grant; see http://emeld.douglass.arizona.edu/. We especially thank Gary Simons of SIL International for his help and advice on all aspects of our project.

References

BATEMAN, J. A., HENSCHEL, R. and RINALDI, F. (2001) The generalized upper model 2.0. http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html

BATEMAN, J. A., MAGNINI, B. and RINALDI, F. (1994), The generalized Italian, German, English upper model, *Proceedings of the ECAI94 Workshop: Comparison of Implemented Ontologies*. (Amsterdam).

BERNERS-LEE, T., HENDLER, J. and LASSILA, O. (2001) The semantic web, *Scientific American*, May 2001.

BICKEL, B. and NICHOLS, J. (2002) Autotypologizing databases and their use in fieldwork, *Proceedings of the Int. LREC Workshop on Resources and Tools in Field Linguistics*. (Las Palmas, 25–26 May 2002).

BYBEE, J. (1985) *Morphology: A Study of the Relation Between Meaning and Form*, Amsterdam/Philadelphia: John Benjamins.

DRYER, M., HASPELMARTH, M., GIL, D., and COMRIE, B. (eds) (forthcoming). *World Atlas of Language Structures*. Oxford: Oxford University Press.

FELLBAUM, C., Ed. (1998) *WordNet: An Electronic Lexical Database*, Cambridge, MA: The MIT Press.

GANGEMI A., GUARINO, N., OLTRAMARI, A. and BORGO, S. (2002) Cleaning-up WordNet's top-level, *Proceedings of the 1st International WordNet Conference* (21–25 January 2002).

GREENBERG, J. (1963) Some universals of grammar with particular reference to the order of meaningful elements, in J. GREENBERG (ed.) *Universals of language*. Cambridge, MA: MIT Press.

MASOLO C., BORGO, S., GANGEMI, A., GUARINO, N., OLTRAMARI, A. and SCHNEIDER, L. (2002) WonderWeb deliverable D17. The WonderWeb library of foundational ontologies. Preliminary Report (ver. 2.0), 15-08-2002.

MONACHESI, P., DIMITRIADIS, A., GOEDEMANS, R., MINEUR, A., and PINTO, M. (2001) The typological database project. Paper presented at the Typological Database Project. (Utrecht, 29–30 June, 2001).

PAYNE, T. E. (1997) *Describing Morphosyntax: A Guide for Field Linguists*, Cambridge: Cambridge University Press.

POLLARD, C. and SAG, I.A. (1987) *Information-based Syntax and Semantics*. CSLI Lecture Notes, Number 13. CSLI.

TALMY, L. (2000) *Toward a Cognitive Semantics. Volume II: Typology and Process in Concept Structuring*, Cambridge, MA: The MIT Press.