Building a Knowledge Base of Morphosyntactic Terminology

William LEWIS
Department of Linguistics
University of Arizona
P.O. Box 210028
Tucson, AZ 85721
wlewis@u.arizona.edu

Scott FARRAR &
D. Terence LANGENDOEN
Department of Linguistics
University of Arizona
farrar@u.arizona.edu
langendt@u.arizona.edu

Abstract

This paper describes the beginning of an effort within the Linguist List's Electronic Metastructure for Endangered Languages Data (E-MELD) project to develop markup recommendations for representing the morphosyntactic structures of the world's endangered languages. Rather than proposing specific markup recommendations as in the Text Encoding Initiative (TEI), we propose to construct an environment for comparing data sets using possibly different markup schemes. The central feature of our proposed environment is an ontology of morphosyntactic terms with multiple inheritance and a variety of relations holding among the terms. We are developing our ontology using the Protégé editor, and are extending an existing upper-level ontology known as SUMO.

Introduction

Our project is a component of the E-MELD project.[1] E-MELD has a number of long-term goals, including creation of:

- Metadata standards for endangered language data;
- Markup recommendations for structuring endangered language data for presentation and analysis on the World Wide Web;
- A "best practices" showcase.

This paper describes the first stage in reaching the second of these goals. Our decision to begin work on the analysis of morphosyntactic terms was based on the recommendations of a markup work group that the Linguist List organized at the Language Digitization Workshop in Santa Barbara, June 21-24, 2001. That group divided the task of developing markup recommendations into several problem areas, and identified morphosyntactic markup as the first problem to be tackled.

1 Linguistic markup

Markup or "tagging" of linguistic structure has existed for as long as language data have been available in machine-readable form. For example, all the words of the Brown corpus (Kucera and Francis 1967) were tagged for part of speech, and the development of more and more comprehensive linguistic tagging schemes has continued to the present time.

1.1 The Text Encoding Initiative

The first and most ambitious effort to develop standards for linguistic markup was the Text Encoding Initiative (TEI), which began work in 1987 and which continues to be very active. The first widely distributed TEI Guidelines (Sperberg-McQueen and Burnard 1994) provide recommendations for various types of linguistic markup using SGML including chapters discussing:

- Speech transcription (chapter 11)
- Print dictionaries (chapter 12)

[1] The E-MELD project is supported by NSF Grant 0094934. The principal investigators are Anthony Aristar, Wayne State University; Helen Aristar Dry, Eastern Michigan University; Steven G. Bird, University of Pennsylvania; Martha S. Ratliff, Wayne State University; and D. Terence Langendoen, University of Arizona.

- Linking, segmentation, and alignment (chapter 14)
- Simple analytic mechanisms (chapter 15)
- Feature structures (chapter 16)
- Feature system declarations (chapter 26)

The last three of these chapters are of particular relevance to our work on morphosyntax. The section on linguistic annotation in chapter 15 defines a set of tags and attributes for segmenting text into sentences, phrases, words, morphemes, and graphemes, and for associating with each of these units other values. It also defines tags and attributes for other types of interpretive markup such as for discourse analysis.

Chapters 16 and 26 together provide methods for building tagsets that represent linguistic structure to any desired degree of detail, and for validating the markup against a grammatical analysis. However, the tags and attributes are designed to represent the formal structure of the analysis only, with the details of the analysis specified as content (Langendoen and Simons 1995).

1.2 Data interchange model

An important goal of the TEI was to provide a means whereby projects in the humanities could exchange and thereby compare their work. For example, suppose two projects have data marked up in accordance with their own local markup scheme, and they wish to compare their analyses, perhaps with an eye to future collaboration. Each project will need to be able to translate its local markup into the TEI interchange format, and to translate documents in the TEI interchange format into its local markup. As a matter of practical functioning, this has required projects to use the TEI interchange format as their local markup, so as to avoid translations at both ends.

1.3 Data comparison model and "metatagging"

The E-MELD project anticipates that users will want to be able to obtain information about endangered languages on the World Wide Web without regard to the tagging schemes that are used in the various websites they consult. Thus it cannot impose a markup standard for endangered language websites, even implicitly by developing a data interchange format. Rather the project itself needs to provide a means to undertake syntactic and semantic analysis of the tagging schemes that the various endangered languages websites employ. In that way, a user who searches for information from those sites can obtain that information regardless of the details of the tagging scheme employed by any particular website.

We define metatagging as providing the resources needed to compare linguistic data sets that use possibly different markup definitions. Metatagging presumes the use of a common markup language across the relevant websites, let us say XML, and provides methods whereby by the semantics as well as the syntax of each tagging scheme can be compared. Here are two simple illustrations of problems that will undoubtedly arise.

- Several sites use an "absolutive" tag. We know from published work in linguistics, that the term can mean "unpossessed form of a noun" or "case of the subject of an intransitive verb or the object of a transitive verb". What does it mean on these websites?
- Some sites use a "possessive" tag, some use "genitive" and a few use both. Again, we know from published work that the terms are sometimes used interchangeably, and sometimes to mean different, but closely related, things. If a user searches for "genitive", should all instances of "possessive" also be returned, or just those in which it means the same thing as "genitive"?

2 The E-MELD Knowledge Base

This section describes the core of the E-MELD knowledge base of linguistic terms. Key to any knowledge base is an ontology, or organized system of concepts that makes explicit what may exist in a domain. A knowledge base is not complete without a logic to provide formal structure and a system of computation to perform useful operations (Sowa 2000). These latter components are still under development and will not be discussed here.

2.1 The Need for an Ontology

A general goal of E-MELD is to allow the end user (the field linguist, the syntactician, the language teacher, etc.) as much freedom as possible with respect to the form of his/her tagset or markup scheme as described in section 2.3. This will facilitate seamless data exchange between scholars not already familiar with each other's work. To this end we have chosen to take a knowledge-based approach which will avoid having to dictate a gold standard of markup terminology. In short, the knowledge base and its accompanying tools will act as an interlingua for data comparison. The system we envision will be able to translate between tagsets and provide a general framework for accessing disparate datasets by using a single search engine format. The key to the system's generality is the ontology. An ontology makes explicit what kinds of concepts exist (in this case linguistic concepts) in a domain; it defines what relations can exist between concepts; and it represents knowledge about the target domain. The goal is to construct as complete an ontology of linguistic terminology as needed for describing all languages in the E-MELD domain.

An ontology can be thought of as an enriched taxonomy. For example, consider the following partial ontology of morphosyntax terminology in Figure 1. The arrows in the taxonomy represent the inheritance or is-a relation. For example the knowledge that a PHRASE is a CONSTITUENT is represented here, as is the knowledge that a NOUNPHRASE is a PHRASE. By virtue of inheritance (which is transitive), the knowledge that a NOUNPHRASE is a CONSTITUENT is also represented. But what about a case where some term stands in an is-a relationship to more than one other term? For example, a PHRASE is a CONSTITUENT and a PHRASE is a CONSTRUCTION by standard definitions. Such a situation warrants the use of multiple inheritance which is illustrated in Figure 2.

Taxonomies, and some ontologies, avoid multiple inheritance. But in any application beyond toy implementations and very simple domains, multiple inheritance is necessary. Whereas a taxonomy only represents is-a relations, an ontology can potentially represent any definable relationship between concepts. For example WORD and PHRASE are in a mereological (part-whole) relationship, namely a PHRASE is composed of WORDS. This may be indicated in the ontology by introducing another link, as in Figure 3.

Other relevant relationships among morphosyntax terms are *expresses*, as in *a SENTENCE expresses a PROPOSITION*, and *modifies*, as in *an ADVERB modifies an VERB.*

Using an ontology to organize terms results in a clean domain model for E-MELD terminology. When fully specified, no ambiguities and no gaps will exist among relations. But more than helping to define the domain, an ontology provides the basis for *logical inference* within the knowledge base. Multiple inheritance is one kind of inference explicitly represented in the ontology fragment in Figure 2. Another kind of inference concerns tags that refer to more than one grammatical relation. For example, consider the hypothetical tag GENS which stands for "genitive-singular". If the end-user is searching for all cases of morphemes that, for example, do not indicate plurality, then GENS by logical inference would be included in the results.

2.2 Choosing an Ontology Editor

Once we made the decision to use a knowledge-based approach and therefore to develop an ontology, we faced several implementational issues. Knowledge-based systems are expensive to construct and maintain, with the most challenging tasks being the construction of the ontology and specification of the representation language. One solution to this problem is to use an off-the-shelf ontology editor to facilitate development of the knowledge base. We chose to use Protégé[2] for several reasons: (1) it offers an extensible architecture for developing knowledge systems; (2) it uses CLIPS[3] style formatting for representing data; (3) it provides JDBC[4] support; (4) it is widely used in the knowledge engineering community; and (5) it is freeware and open source.

---

[2] Protégé is a product Stanford Medical Informatics and is available at http://protege.stanford.edu/.

[3] See http://www.ghg.net/clips/CLIPS.html for an overview of the CLIPS system.

[4] JDBC stands for *Java Database Connectivity*. See http://java.sun.com/products/jdbc

Figure 1 - Partial Ontology of Morphosyntactic Terminology

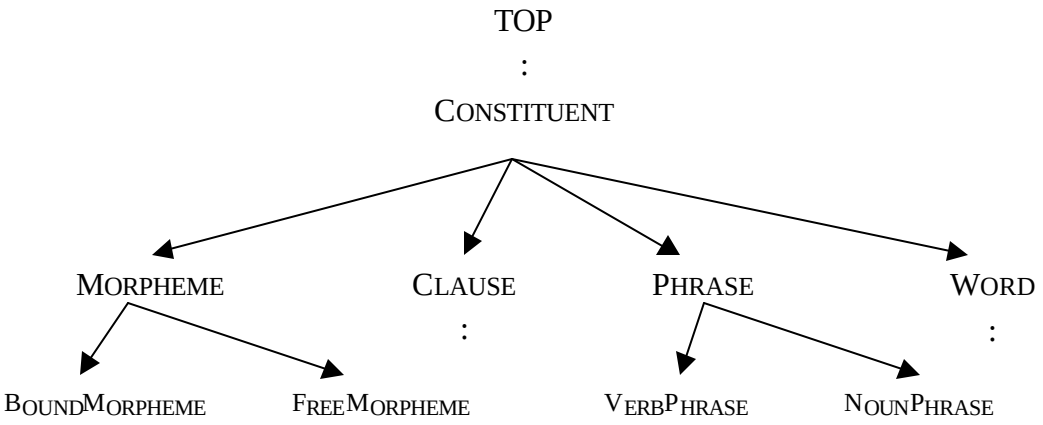


Figure 2 - Multiple Inheritance

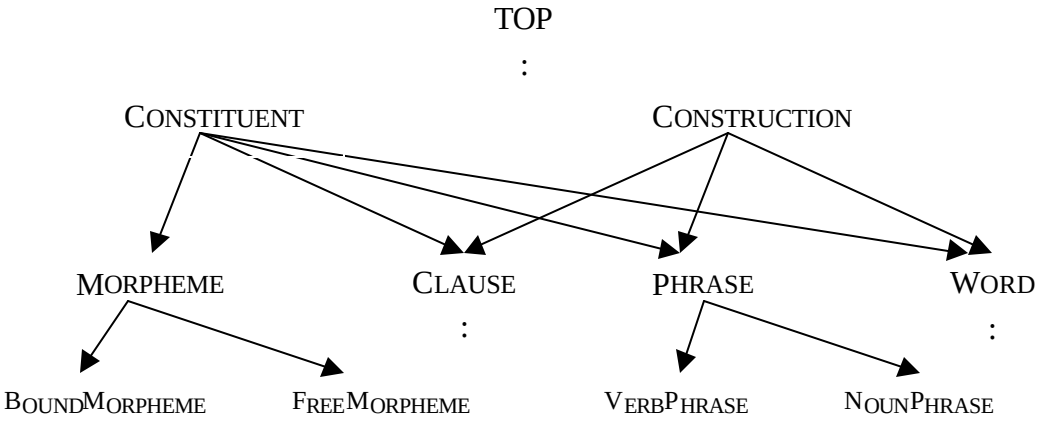
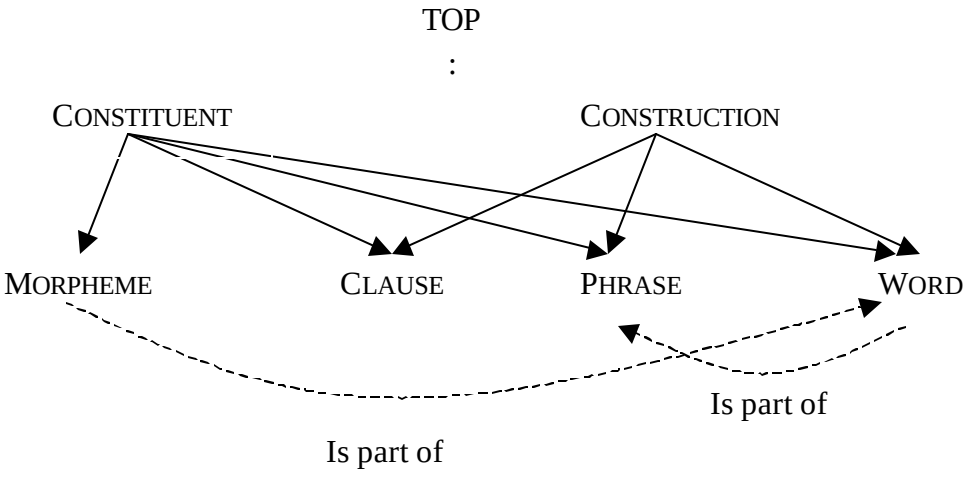


Figure 3 - Ontology with Mereological Relation Added



As mentioned earlier in this section, Protégé uses the standard CLIPS representation language for representing an ontology. The three most important types of data structures in CLIPS are classes, slots, and instances. Classes are the core of the ontology. Essentially, an ontology is a hierarchy of classes as in Figures 1-3. Thus, a class indicates a set of concepts that can exist in the domain. Slots are a way to represent relations among classes. For example, the class SYMBOLICSTRING has slot REPRESENTS which has as an argument GRAMMATICALPROPERTY. The slot indicated that a SYMBOLICSTRING represents a GRAMMATICALPROPERTY. An instance represents an actually entity that exists, for example the orthographic tag GEN is in an instance of the abstract class Tag.

2.3 Enriching an Existing Ontology

Protégé facilitates building an ontology. But even with an ontology editor, starting from scratch is very time consuming and difficult. In order to save time and effort we decided to augment an existing ontology called SUMO (Suggested Upper Merged Ontology). SUMO was created by the Teknowledge Corporation[5] with extensive input from the knowledge engineering community, in particular from the IEEE SUO (Standard Upper Ontology) Working Group[6] (Niles and Pease 2001). The ultimate goal of the working group is to have a standard upper ontology that can be used for any application.

Using SUMO offers several advantages. The primary one is that since SUMO is an upper ontology the top level categories are already defined. Furthermore, SUMO represents several years of research and development in the field of knowledge engineering and is considered a state-of-the-art effort. It is envisioned that as later versions of SUMO are released the E-MELD ontology will benefit from its enrichment. Finally, since SUMO is used in a wide variety of knowledge-based applications, the possibility exists for using these in expanding the E-MELD effort to handle other linguistic challenges as they arise.

[5] http://ontology.teknowledge.com/cgi-bin/cvsweb.cgi/SUO/

[6] http://suo.ieee.org/

2.4 Considerations for an Ontology of Linguistic Terminology

Once the ontology is fully specified, it will be possible to represent and reason about any dataset that a researcher submits. Only minimal changes will be needed to maintain the knowledge base for use with each new dataset. In developing the ontology our first step was to collect a comprehensive inventory of morphosyntactic terminology. Two sources we consulted were a morphosyntactic glossary developed by SIL International[7] and the DOBES[8] (Dokumentation Bedrohter Sprachen) list of morphosyntactic terms. In addition to the terminology gleaned from these sources, we also relied on a number of texts on linguistic terminology (Crystal 1997; Payne 1997) and several dictionaries and grammars of endangered languages (Hill et al 1998; Dedrick and Casad 1999). Once a substantial body of terminology was created, we categorized each term and placed it in the ontology as either a class or an instance of a class.

Within the domain, we have identified three major categories of concepts to include in the ontology. First is the category subsuming *linguistic objects and processes*. For example, WORDS, PHRASES, MORPHEMES, and CLAUSES are linguistic objects referring to mental representations of their corresponding phonological or written forms. Likewise, some processes include SUFFIXATION, METATHESIS, and SYLLABIFICATION. Second is the category subsuming what may be called *grammatical properties* of the first category. Examples of these include CASE, TENSE, NUMBER, and ASPECT. Third is the category subsuming all possible tags a researcher wishes to use to describe a given language, for example PROG for progressive, NEG for negative, DIM for diminutive, and EX for extreme (Hill et al. 1998). This category consists of orthographic tokens. Each token represents either a linguistic expression or a grammatical property. The key to constructing the knowledge base is linking each of the researcher's collection of tags to elements in the ontology, i.e., the first two categories.

[7] http://www.sil.org/linguistics/glossary/

[8] http://www.mpi.nl/DOBES

2.5 The E-MELD System Architecture

The architecture for the envisioned system is given in Figure 4. The three major components of the E-MELD system are (1) the graphical user interface (GUI), (2) the knowledge base (containing the ontology and query engine), and (3) the database of endangered languages marked up in XML format. The end user will be able to access the E-MELD system via the World Wide Web as the knowledge base and language data will reside together at a remote site. The user may pose queries to the knowledge base in standard search engine format (similar to that of Yahoo or Google). For example, the query "ergative P2" will return a list of languages and/or actual language data from P2 languages containing ergative constructions. The only requirement that is required is that the documents containing the individual language data be in XML format. The query engine will have access to XML metadata and all language data in each file. Once the envisioned system is implemented only minimal maintenance will be required to add additional language data. Adding new data sets merely requires the ontology manager to interpret the researcher's tagset and to incorporate it into the existing ontology.

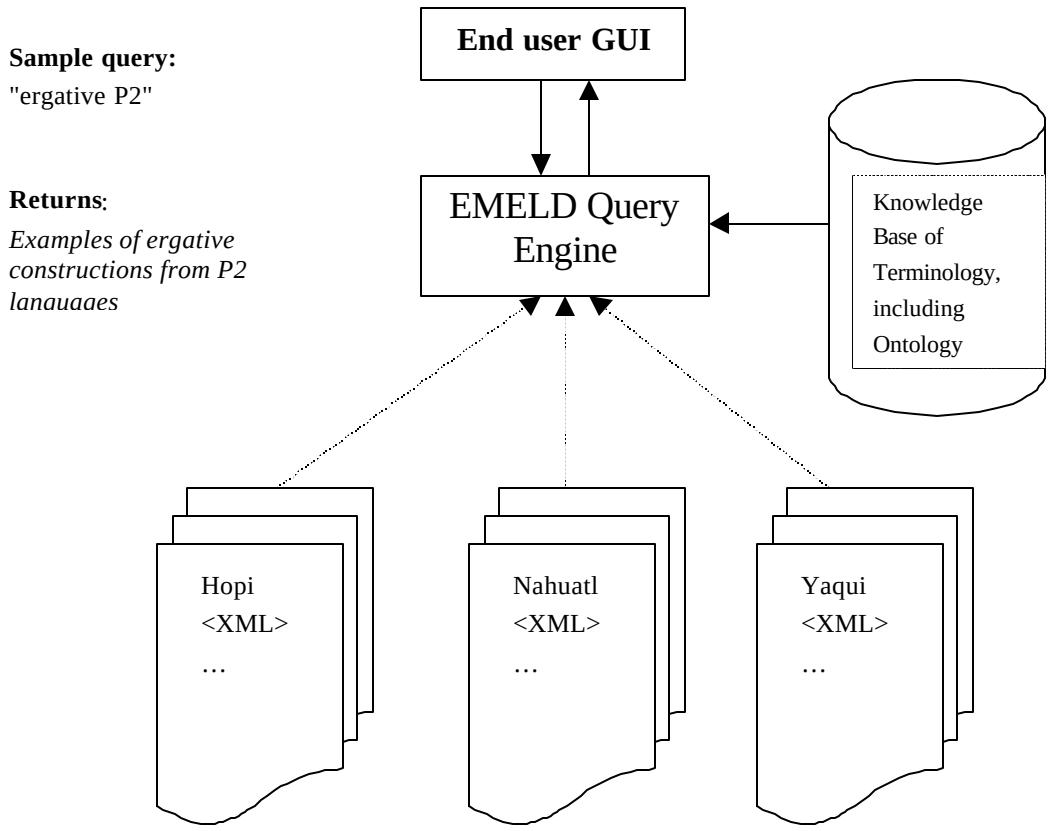
3 Future Directions

Revision to the existing E-MELD ontology—now consisting of over 1,000 entries—continues. In addition to revising the ontology following the strategies used so far, four additional directions are foreseen:

(1) Subject the existing morphosyntactic ontology to peer review: At the Language Digitization Workshop mentioned in section 1, a panel of researchers consisting of field linguists, computational linguists, corpus linguists, and theoretical linguists was assembled to launch E-MELD's markup effort. In January 2002, the first phase of the project will be complete, and the ontology will be submitted to this panel for review and revision.

(2) Merge existing tagsets into the ontology: We

- **Figure 4 - E-MELD Knowledge Base and Query Architecture**



are currently incorporating the tagset used in the Hopi Dictionary (Hill et al 1998) into the ontology.

(3) Extend the ontology beyond the morpho-syntactic domain: We are currently reviewing resources for adding phonological and discourse level/pragmatic elements to the ontology.

(4) Develop tools that use the ontology: The intended goal of the ontology is to provide a standardized means for linguistic data comparison. Towards that end, database and query tools will be developed that use the ontology. Because the development of such tools from scratch can be quite time consuming, existing tools and resources will be used and adapted to whatever extent possible.

References

Crystal D. (1997) *A Dictionary of Linguistics and Phonetics*, 4th Edition. Oxford: Blackwell.

Dedrick J. M. and Casad E. H. (1999) *Sonora Yaqui Language Structures*. Tucson: University of Arizona Press.

Hill, K. C., Sekaquaptewa E., Black M. E., Malotki E., and Lomatuway'ma M. (1998) *Hopi Dictionary = Hopìikwa Lavàytutuveni: a Hopi-English Dictionary of the Third Mesa Dialect with an English-Hopi Finder List and a Sketch of Hopi Grammar*. Tucson: University of Arizona Press.

Kucera H. and Francis W. N. (1967) *Computational Analysis of Present-day American English*. Providence: Brown University Press.

Langendoen D. T. and Simons G. F. (1995) A rationale for the TEI recommendations for feature-structure markup. *Computers and the Humanities* 29.191-209.

Niles I. and Pease A. (2001) Towards a Standard Upper Ontology. In C. Welty and B. Smith (Eds.) *Formal Ontology in Information Systems: Collected Papers from the Second International Conference*. New York: ACM Press, pp. 2-9.

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

Sowa J. (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks/Cole.

Sperberg-McQueen C. M. and Burnard L. (1994) *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Chicago and Oxford: Text Encoding Initiative.