A Common Ontology for Linguistic Concepts

Scott Farrar, William D. Lewis, and D. Terence Langendoen
{farrar, wlewis, langendt}@u.arizona.edu

Abstract

As part of a project called Electronic Metastructure for Endangered Languages Data[1] (EMELD), we have developed an ontology of concepts that encompasses a wide range of linguistic phenomena. The idea was initially conceived to facilitate both the knowledge sharing of annotated linguistic data and the searching of disparate language corpora. Such an ontology, however, is needed outside of the EMELD project for enhancing performance of the Semantic Web, for developing expert systems capable of linguistic analysis, and for providing a theory-neutral backbone in the processing of scientific documents pertaining to the linguistics domain. With an eye toward acceptance by the knowledge engineering community in general, we built the linguistic ontology on top of the Standard Upper Merged Ontology (SUMO).

1 Introduction

During the past century hundreds of the world's languages have become extinct. Of the roughly 6,000 remaining languages, half are in danger of extinction during the first half of this century. Preserving the record of language that has disappeared provides native communities with a link to their culture and ensures that language researchers will have access to data for linguistic analysis. To this end, EMELD (Electronic Metastructure for Endangered Languages Data) has as one of its mandates to develop software that facilitates storage and sharing of endangered language data. Data on endangered languages exists in many forms, from historical field notes, some dating back hundreds of years, to modern sound recordings. Typically, a linguist will go into the field and collect word lists, transcribe speech, or record interviews. The result is a *linguistic dataset* (a grammar, dictionary, or glossed corpus) of the language in question. In the ideal case, the author of the dataset will employ what is called *markup* to annotate certain features of the language important for scientific study. It is of great scientific interest to bring together resources from languages like Hopi, Mocovi, and Biao Min in order to make them available to search and retrieval on the emerging Semantic Web (Berners-Lee, Hendler, and Lassila 2001). The result will not only help to preserve cultural heritage, but will also enable broad access to a disappearing body of scientific data. One major obstacle to this endeavor is that datasets are almost always incompatible on a number of levels, the *data interoperability problem.* A solution would be to dictate standards for the form and content of linguistic markup. However, we argue that this approach is both impractical and unnecessary and most critically will not be accepted by the linguistic community. Instead we are developing a knowledge rich system which utilizes *linguistic metadata* (a description of linguistic data) to solve the problem of data interoperability. The most important component of the system is the linguistic ontology which we developed on top of the existing Suggested Upper Merged Ontology (SUMO) (see Section 4.1).

[1] Supported by NSF Grant 0094934.

2 The Problems

The difficulties that we face with the search and retrieval of linguistic datasets are common to the database community in general. These include the problems of incompatible data formats and managing dynamic information (Singh 1998). Further, linguistic datasets can vary according to markup scheme, natural language semantics, and theoretical style and, furthermore, can change over time as new data is gathered.

2.1 Incompatibility of Datasets

Linguistic datasets may vary according to markup scheme. Even though two annotators are working within the same theory, they may still disagree on certain notational conventions. In the simplest case, field worker A may use PFV to annotate a morpheme expressing the perfective aspect, while field worker B may use PERF. Our solution is to link the two notations via the well-defined concept PerfectiveAspect[2]. A more complicated example involves cases where morphemes are labeled for one concept but really conflate two or more, as in Hill et al.'s (1998) markup of Hopi. The morpheme labeled CAUS is actually causative and perfective. Another language researcher would have trouble discovering this fact while, say, searching for all instances of the perfective in Uto-Aztecan languages. Given the ontology and the necessary assertions, "Hopi CAUS implies PERF", the correct inference could be drawn in an automated search.

Secondly, datasets can vary according to natural language semantics. The problem can be illustrated by a query that is meant to return instances of the relation expressed by the English preposition *into*. Consider the way Hungarian and English express directionality, a thematic relation describing space and motion that holds between a predicate and its arguments. Whereas Hungarian expresses movement by a complex case system, English communicates the same notions via its inventory of prepositions, as in (1):

(1) az épület-be mégy-ek
the building-Illative go-1P/Sg
I am going into the building.

Linguistic concepts, then, are grounded in language-neutral concepts, in this case spatio-kinetic ones. We have built the linguistic ontology on top of the Standard Upper Merged Ontology (SUMO). Our work may be viewed as an extension of SUMO into the domain of linguistics. We anticipate that our ontology could be incorporated into any standard upper ontology, but SUMO is our initial test case.

A third problem is that researchers may use different criteria (often but not necessarily theory-based) which affect how a linguistic concept is analyzed. We give two examples. First, the concept of "action habitually (or customarily) undertaken by the subject" may be expressed by a verbal affix. Notionally, this concept is a type of verbal aspect, analogous to such clearly aspectual categories as RepetitiveAction. So understood, it should be able to combine with other affixes expressing Tense (time relative to the time of speaking). However in some languages, such as Hopi, it contrasts morphologically with tense morphemes, and so is analyzed in such languages as HabitualTense, rather than as HabitualAspect. Second, in one linguistic theory, every noun and pronoun in every language is marked for Case, so that in the English sentence "Tyson bit Holyfield", "Tyson" is marked for NominativeCase and "Holyfield" for AccusativeCase. In another, only those nouns and pronouns are marked for Case when it is realized overtly, so that in English the pronouns "he" and "him" are marked for case, but not the nouns "Tyson" and "Holyfield". The solution here is to distinguish two notions of Case, one for each theory, for example

[2] Concepts are capitalized.

MorphologicalCase and AbstractCase.

Finally, many of the datasets, we anticipate, will be dynamic. Field workers who gather linguistic data are constantly adding to and revising their work. It is unrealistic to expect that a dataset, once constructed, will remain static, especially considering that new opportunities will arise for collection and that the authors will want to keep pace with current linguistic theory. This presents a challenge to the archive aspect of EMELD. How can we ensure the data will not change to a degree that renders interoperability impossible? Keeping pace with changing data often requires expensive middleware that must be constantly updated (Singh 1998). But taking the data out of the hands of its authors is not feasible either. First of all authors will be reluctant to give up their data if they will no longer have access to it (cite Santa Barbara conference). And if data needs to be updated, it would be the responsibility of the archive. Finally, a separation of the author and data would eliminate the specialized knowledge that the author could contribute.

2.2 Standardization is not Viable

Of course a possible solution to data interoperability is always standardization. For the EMELD project, markup standards would need to encompass both form and content, as per the discussion in Section 2.1, and address language specific semantics and analyses based on a specific theory. Standardization in linguistics is met with the same skepticism as in other fields. In certain limited sub-domains, standardization may be desirable. 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), linking, segmentation, and alignment (chapter 14), simple analytic mechanisms (chapter 15), feature structures (chapter 16), and 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).This effort, though laudable, is inadequate for our purposes. the simple analytic mechanisms are just that, too simple, and the feature structure proposals have too much encoding overhead to be practical for most field linguists and the linguistic community.

As another example of standardization in action, the EAGLES's Corpus Encoding Standard (CES) (Ide and Romary 2000) have implemented minimal content standards for language corpora. The XML Corpus Encoding Standard is part of the guidelines and provides a suit of DTDs for encoding basic document structure and linguistic annotation. The EAGLES standard is comparable to the TEI simple analytic mechanism scheme. Moreover, it is based on the linguistic properties of the languages of Europe and so are insufficient for that reason as well.

Consequently, while standardization efforts such as the TEI and the CES provide for "best-practice" methodology, the majority of

the linguists still prefer their own standards and will no doubt be reluctant to adopt any form of standardization.

3 Towards a Solution

EMELD, then, is a balancing act between the need for data consistency on the one hand and the impracticality of standardization on the other. The following explains how we chose the data model and describes the architecture of the EMELD system.

3.1 Choosing a Data Storage and Distribution Model

In order to have access to such a wide variety of linguistic data, EMELD is designing a data storage infrastructure that provides for interoperability among data sources and gives maximum control to the individual authors. Our approach is a hybrid of what we call the *strong* and *weak* models of data storage and distribution.

The strong model emphasizes centrality and standardization. All linguistic data sets would be stored together in one archive much like the model used by the Linguistic Data Consortium[3]. As a result, all datasets must remain static or be maintained by the central archiving body. The major objection within the endangered languages community is that control over individual datasets would pass out of the hands of its authors, which would not encourage authors to submit their data. Nevertheless, the strong database model provides a number of advantages. First of all, all the datasets could be converted into a uniform format. Data uniformity would allow generalized queries, and the entire archive could be searched with the same tools. The strong database model allows for speed of search and consistency of results. The downside is converting all the incoming data into a format consistent with all the other datasets in the archive, requiring a huge amount of work on the part of the archive team and would require them to deviate from what the original authors had in mind for their data.

The primary advantage of the weak model is that it gives maximum control to the authors of linguistic datasets. This encourages anyone with data to make it available to the EMELD community. Individual authors would be able to maintain their own datasets and use markup that reflected the needs of the particular object language and their own theoretical style. EMELD would mine the data periodically for updates and revisions, much like existing search engine on the World Wide Web. As with the Web, freshness of data would always be an issue. Furthermore, we expect that the technical expertise of contributors will vary widely, resulting in a disparity of the quality and possibly the amount of data that can be maintained. The biggest disadvantage, though, would be in maintaining data consistency, a crucial requirement for a generalized query engine. The problem would only be compounded as the number and variety of datasets increased.

After evaluating these two models, we decided to implement a hybrid approach that takes advantage of both of the above models while eliminating their biggest liabilities. The centerpiece of the hybrid approach is an ontology of linguistic concepts. The ontology is a broad coverage knowledge base of linguistic concepts with a precise semantics. By mapping the linguistic annotation schemes used in the archived datasets into the ontology, we essentially eliminate the most serious problem associated with such a massive archiving task, that of standardization. Authors will more readily make their data available if they have maximum control over their own work. With the hybrid approach, EMELD maintains sites for a particular author who supplies the data. To ensure that the archive

[3] The LDC is an large archive that provides tools for sharing and preserving language data (http://www.ldc.upenn.edu/).

is kept up to date, this data is mined periodically for updates.

The only standard that the EMELD project will impose on the datasets is they be marked up in XML. To facilitate the conversion to XML, our team is also responsible for developing a suite of tools to aid field linguists and other researchers to record and markup their data. Once the data is marked up, a one-time mapping to the linguistic ontology will be carried out. This will ensure compatibility with existing markup, semantic interpretations, and theoretical styles. (We are further considering recommendations with include an RDF specification.)

3.2 Overview of the EMELD System

The datasets (from endangered languages like Hopi, Mocovi, and Biao Min) will represent a small slice of the Semantic Web. The EMELD system will be the access portal which will allow for smart searches of the endangered language data. The four major components of the EMELD system are (1) the graphical user interface (GUI), (2) the query engine, (3) the linguistic ontology, and (4) the datasets of endangered languages. The end user will be able to access the EMELD system via the Semantic Web. The resulting EMELD system is given in Figure 1.

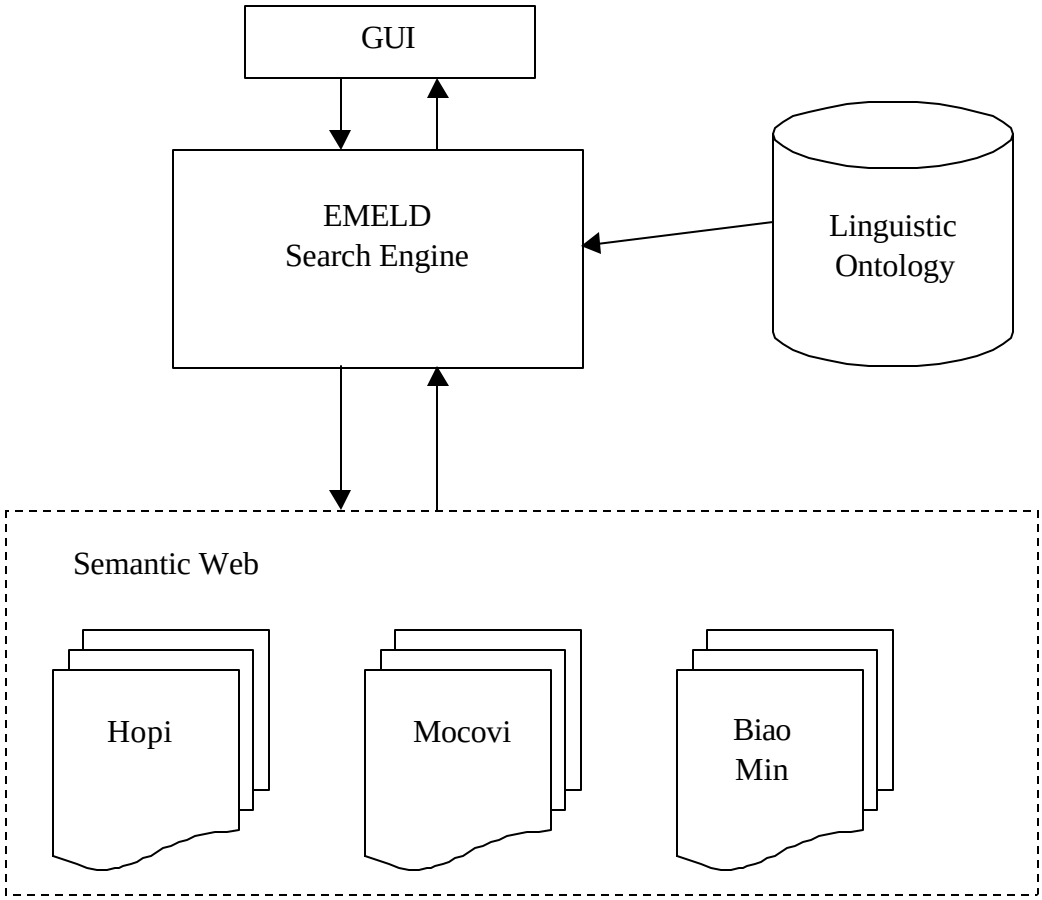


Figure 1: Process Model for the EMELD System

4 The Linguistic Ontology

The linguistic ontology contains what can be considered *metadata*, which is commonly described as data about the data (Tannenbaum 2002). However, the ontology is really much more. It is a repository of linguistic knowledge that attempts to ground linguistic constructs in concepts of time space, causality and human interaction.

4.1 Conceptual Modeling of the Linguistics Domain

We use language to talk about our world. We talk about time, space, causality, and other events relevant to human interaction. We talk about specialized scientific concepts, birthday parties, numbers, and gods. The domain of linguistics could be construed to cover everything under the sun that we talk about. Fortunately, in order to establish data interoperability between datasets of endangered languages, we can limit the domain of linguistics to cover only those topics relevant to grammar, albeit grammar as it is broadly interpreted. Currently our ontology includes the base concepts pertaining to linguistics and language, including Grammar, Language, Dialect, LinguisticExpression, and LanguageFamily. The majority of the work that is relevant to metadata has been carried out in the sub-domain of morphosyntax. By morphosyntax we refer "to grammatical categories or properties for whose definition criteria of morphology and syntax both apply, as in describing the characteristics of words" (Crystal). Because linguistics encompasses concepts from so many different domains, we chose to build the linguistic ontology on top of the Suggested Upper Merged Ontology (SUMO). SUMO (Niles and Pease 2001) is an existing framework specifically designed to provide a basis for more specific domain ontologies. We chose SUMO for a number of reasons: it combines a number of top-level ontologies to achieve wide conceptual coverage, it had a strong basis of semiotic and linguistic concepts already (including those derived from the verb classes of Levin 1993), and it is being developed by an IEEE working group that includes a number of experts from a variety of fields.

```
Entity
  Physical
    Object
      ContentBearingObject
        Icon
        SymbolicString
        LinguisticExpression
          WrittenLinguisticExpression
            Text
            Sentence
            Phrase
            Word
            Morpheme
          SpokenLinguisticExpression
            Dialogue
            Sentence
            Phrase
            Word
            Morpheme

    Process
      Communication
        Stating
        Directing
        Declaring

  Abstract
    Class
      Relation
        Predicate
          GrammaticalRelation
            Aspect
            Tense
            Case
            Agreement

    Attribute
      GrammaticalAttribute
        Gender
        Person
        Number
```

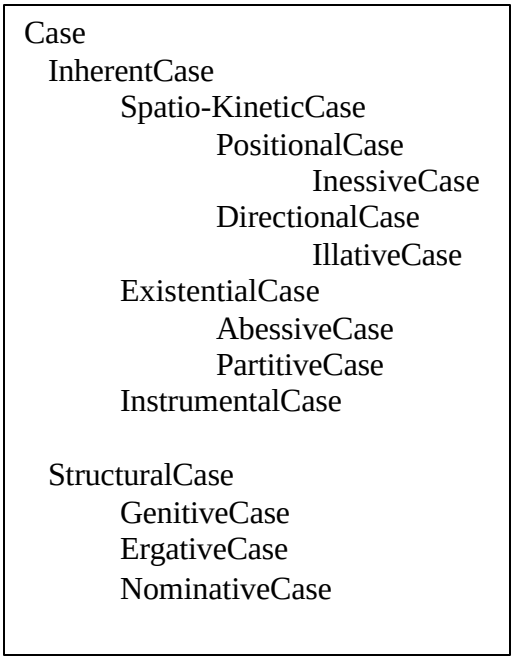
Figure 2: Linguistically Relevant Concepts in Relation to the SUMO Top Level

4.2 Details of the Ontology

As much as possible we tried to use existing elements of the SUMO. First of all SUMO already includes a good semiotics architecture for the representation and the communication of information in general. Expanded from the original SUMO somewhat are the basic segments of language, which are classified as LinguisticExpressions: Text, Sentence, Phrase, Word, Morpheme, and PhonemeSequence. There is a further distinction in the form of the LinguisticExpression which can be Written, Spoken, or Signed. We argue that a WrittenLinguisticExpression is a fundamentally different concept than a SpokenLinguisticExpression, and not simply an attribute like male/female. The distinction is particularly important when reasoning about the *equivalentContentRelation*[4] between, for example, what someone says and what is printed in a newspaper. There are a number of other semiotic predicates that apply, but some of the most important are *refers*, *represents* and *containsInformation*. *Refers* is the general process of semiosis whereas *represents* is used in the sense that some entity expresses, connotes or describes something else. Finally, *containsInformation* is more linguistically specialized, as it relates the ContentBearingObjects to actual Propositions.

Already present in the SUMO structural ontology is the concept of a CaseRole, which covers the familiar linguistic notions of agent, patient, and theme. Concerning the markup of endangered languages, we anticipate that many datasets will use these CaseRoles or their equivalents. Also, the concepts of Number, PropositionalAttitude and the basic notion of a Predicates is already in SUMO. It is worth mentioning here that the concept Language itself has been controversial and the details have not been fully worked out. We argue at least for the subclasses NaturalLanguage, ComputerLanguage, and AnimalLanguage.

For our primary concern, the domain of morphosyntax, there are the concepts of Case, Aspect, Tense, and GrammaticalRelations. These concepts were not in the SUMO and make up a large part of our conceptual modeling. We propose the inclusion of a subclass GrammaticalRelation as a subclass of Predicate (which subsumes other relational categories: SpatialRelation, TemporalRelation, etc.). The modeling of Case is a good example of the far reaching import of linguistic concepts. As mentioned in Section 2.1, Case can specify a particular relation between a predicate and its arguments. Many case systems in natural language relate to the spatio-kinetic relations of an action to its direct or indirect participants. Hungarian-type languages in particular have rich spatio-kinetic case systems (roughly equivalent to prepositions English-type languages). Our proposal for the basic organization of Case is shown in Figure 3.



[4] Predicates are italicized.

Figure 3: Morphosyntactic Case, its Subclasses, and Instances

Case as represented in Figure 3 overlaps somewhat with those concepts subsumed by CaseRole in (2). One example is InstrumentalCase. The instrument is a particular CaseRole but is marked in the morphology by the instance of Case.

One solution to the interoperability problem is to map Hungarian cases directly to English prepositions. But this would require a many-to-many mapping between all languages in question, which we argue is too difficult and furthermore unnecessary. Instead, the relevant Hungarian cases and English prepositions are defined in terms of more general spatial concepts of motion in the ontology. For this particular example, the concept described by *–b(V)* or *into* is defined by using well-defined categories and predicates already in the ontology as in Figure 4.

```
(=>    (?X Object)
       (?Y Object)
       (outside ?X ?Y)
       (?M Motion)
       (patient ?M ?X)
       (inside ?X ?Y))
```

Figure 4: Concept of "Motion Into" Expressed in SUO-KIF

Finally, in laying out the backbone taxonomy we employed recommendations laid out in Guarino and Welty (2002). One example in particular deserves notice. Guarino and Welty distinguish between rigid, semi-rigid, and anti-rigid properties. Rigid properties are essential to all instances of a subclass. While anti-rigid refers to the non-essential ones. The notion of a rigid property correlates with GrammaticalAttributes of Gender, Number, and Person.

4.3 Evaluating the Ontology

Our ontology is available on the EMELD website (http://emeld.douglass.arizona.edu), and we invite commentary on all of its contents. We have posted the discussion of the more general concepts that pertain to language and linguistics to the SUO discussion group. Beyond the possibility of peer critique, the ontology is evaluated every time we incorporate a new dataset. Each dataset, which usually encompasses a new language, is compared against the existing model. This implies that the ontology is data driven. That is, no new categories are added unless the data specifically warrants it.

5 Conclusion

Although we have constructed the ontology and its architecture is firmly established, the EMELD system is still in the beginning stages. Remaining tasks include developing the search engine and tools for encoding the endangered language data. Future directions for conceptual modeling include the domains of phonology and discourse analysis. Further, the linguistics ontology has applications beyond the immediate EMELD project, in that it can be used as part of an expert system for reasoning about language data or as part of an interlingua designed for machine translation systems.

References

Berners-Lee, T., Hendler, J., and Lassila, O. (2001) The Semantic Web. *Scientific American*, May 2001.

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

Guarino, N. and Welty, C. (2002) Evaluating Ontological Decisions with Ontoclean. *Communications of the ACM*, (45), pp. 61-65.

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.

Ide, N. and Romary, L. (2000) XML Support for Annotated Language Resources. Linguistic Exploration Workshop on Web-Based Language Documentation and Description, Dec. 12-15, University of Pennsylvania.

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.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press.

Niles, I., & Pease, A., (2001), Toward a Standard Upper Ontology, *in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).*

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

Singh, N. 1998 Unifying Heterogeneous Information Models. *Communications of the ACM*, Volume 41, (5): 37-44.

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

Tannenbaum, A. (2002) *Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand*. Boston: Addison-Wesley.