Bridging the Markup Gap: Smart Search Engines for Language Researchers

D. Terence Langendoen, Scott Farrar, and William Lewis

Department of Linguistics, University of Arizona

{langendt, farrar, wlewis}@u.arizona.edu

This paper describes the design features of a search engine for linguistic data being developed at the University of Arizona as part of the Electronic Metastructure for Endangered Languages Data (EMELD) project. It is intended to access websites in which linguistic data is encoded in RDF/XML, identifies the relevant features, links them to an ontology of linguistic terminology, and supports queries over the terms found both in the ontology and the websites.

Introduction

This paper reports on work in progress being carried out by the University of Arizona component of the Electronic Metastructure for Endangered Languages Data (EMELD-Arizona) project since the publication of Lewis, Farrar and Langendoen (2001). In that paper, we described our effort as an attempt "to construct an environment for comparing [endangered languages] data sets [posted on the Internet that use] possibly different markup schemes" (p. 150). We also stated that we would initially limit the scope of our effort to information about the morphosyntactic properties of those languages.

Varieties of encoding of grammatical information

Without imposing any specific requirements on data providers as to how they encode grammatical information, we can expect to find it to be encoded sometimes in entity names as in (1); sometimes in attributes as in (2); and sometimes in content as in (3). The problem is independent of whether the site uses 'inline' or 'standoff' markup. For ease of exposition, we assume inline markup in this discussion.

(1) `<noun>lirrppi</noun>`

(2) `<word POS="noun">lirrppi</word>`

(3) `<word><feature>POS</feature><value>noun</value><content>lirrppi</content></word>`

In order to facilitate our handling of data in such variety of forms, we would expect to find metadata statements about the way the data are encoded, and information about what abbreviations represent (e.g. that `POS` in (2) represents "part of speech"). Otherwise we would have to contact the site managers for help, and failing that to use our best judgment.

Once we have established an interpretation of a site's encoding of morphosyntactic information, we need next to determine its relations to the interpretations of other sites' encodings. To do this, we proposed to construct an ontology, or knowledge base, of morphosyntactic terminology that would be part of an interface in which users could query or search the sites to which the ontology is linked. In this paper we expand our discussion of the EMELD-Arizona system and instantiate it as a part of the emerging Semantic Web (Berners-Lee, Hendler, and Lassila 2000). We envision the Semantic Web as a potentially ideal environment for bringing the endangered language communities together through the sharing of data. We discuss our component in terms of the two key

enabling technologies of the Semantic Web, Extensible Markup Language (XML) and the Resource Description Framework (RDF, Lassila and Swick 1999).

Querying the data in electronically encoded interlinear glossed examples

Morphosyntactic information can be presented in a variety of ways electronically, in the form of grammatical descriptions, in lexicons and dictionaries, and in glossed examples and texts. We consider here glossed examples, which have appeared in print documents for nearly a century, and which (along with glossed texts) can now be constructed electronically using a variety of software packages. A typical glossed example is the Warumungu example in (4), from Simpson (1998: 727). In (5), we provide an XML encoding of the information represented in (4). The boldfaced material in (5) is discussed below in connection with example (7).

(4) 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.'

(5)
```
<example ref="simpson022">
    <word>
        <morpheme type="stem">
            <spelling>yama</spelling>
            <gloss>leave</gloss>
        </morpheme>
        <morpheme type="clitic">
            <spelling>ajurnu</spelling>
            <gram>3PL</gram>
            <gram>NS</gram>
        </morpheme>
    </word>
    <word>
        <morpheme type="stem">
            <spelling>jarti</spelling>
            <gloss>other</gloss>
        </morpheme>
        <morpheme type="affix">
            <spelling>ki</spelling>
            <gram>DAT</gram>
        </morpheme>
        <morpheme type="clitic">
            <spelling>karn</spelling>
            <gloss>now</gloss>
        </morpheme>
    </word>
    <word>
        <morpheme type="stem">
            <spelling>pikapikka</spelling>
            <gloss>children</gloss>
        </morpheme>
        <morpheme type="affix">
            <spelling>ka</spelling>
            <gram>DAT</gram>
        </morpheme>
    </word>
```

```
        <word>
            <morpheme type="stem">
                <spelling>ngu</spelling>
                <gloss>lie</gloss>
            </morpheme>
            <morpheme type="affix">
                <spelling>nngara</spelling>
                    <gram>OPT</gram>
                    <gram>FUT</gram>
            </morpheme>
        </word>
        <translation>Some should be left for the other
children.</translation>
    </example>
```

From the content of the `gram` elements, and from the list of "annotated abbreviations" in Simpson (1998: 732-734), we can determine that Warumungu contains segments that represent such morphosyntactic structure as NONSUBJECT (a property of "pronominal clusters"), DATIVE case, FUTURE verbal inflection, and OPTATIVE verbal inflection. Neither Simpson nor the editors of the volume in which her article appears say what is meant by 3PL, but we assume that it means what linguists generally use it to mean, namely the combination of THIRD person and PLURAL number.

If we limit ourselves to queries within the set of Simpson's Warumungu examples we do not need to reconcile her use of grammatical terminology with anyone else's. The fact that she defines OPTATIVE simply as a verbal inflection, and not more precisely as a verbal mood, does not prevent us from inquiring into the use of this information in her data. On the other hand, we may wonder whether her use of OPTATIVE is comparable to the use of that term in the analysis of another language in which it is explicitly defined as a mood operator indicating a desire or wish on the part of the speaker or subject. Suppose that we have Simpson's data set and another in which OPTATIVE is so defined. Should a query asking for examples of optative mood return examples from both data sets? The answer here is presumably yes, although it may be useful for the system to flag the examples like (4) as possibly not conforming to the customary definition of OPTATIVE.

Simpson's use of DATIVE case raises other questions. She defines it as either a grammatical case of objects or a semantic case of indirectly affected participants or directions (1998: 733), and from her discussion of (4), we learn that its occurrence there is as a grammatical case: *yama* assigns DATIVE case to its object. However, the gloss does not explicitly indicate this, so a query for a listing of examples of grammatical (as opposed to semantic) DATIVE case cannot be answered. We know, however, that the distinction Simpson makes is found in other languages, so that in the ontology, DATIVE case can be defined generally as compatible with both the grammatical and semantic functions found in Warumungu (and presumably with others as well), with the more specific functions defined as hyponyms. Should someone then wish to refine the markup of examples like (4) to indicate that grammatical as opposed to semantic DATIVE case is present, the new tag can be linked to the appropriate hyponym in the ontology.

The linguistic ontology

The previous section motivated a need for a precise semantics of linguistic metadata. The linguistic ontology is the resource that makes the meaning of the metadata explicit. The

first requirement for meaning is that entities in the domain be defined. Simpson's use of DATIVE illustrates the need for the ontology. First of all DATIVE assumes the existence of the concept case. With case, there are case assigners (e.g., verbs) and case assignees (nouns). Furthermore, case is indicated by case markers (e.g., suffixes). It is possible to speak of the predicate assignsCase. And we can talk about how the verbs, nouns, and case morphemes make up a sentence and that sentences compose to form discourses. Thus, the entities of the domain begin to emerge. Representing the entities of a domain is facilitated by a formal ontology. Together with other facts about the domain, a knowledge base of linguistic terminology can be constructed. An ontology, a well-defined logic, and a means of computation are the foundation of modern expert systems (Sowa 2000). Expert systems are able to reason about specific domains and solve problems. In this case we are interested in solving the problem of data interoperability. If two different researchers, constrained by different theories or the needs of different languages, use conflicting tagging schemes, then an expert system, like a linguist, would be able to decipher and reconcile both data sets.

The ontology lays out what may be discussed and reasoned about. Within the domain of EMELD, there are three types of entities present in the ontology. First of all there are the linguistic concepts themselves, like NOUN or SENTENCE , that make up the core linguistic ontology. There are also the labels, or tags, (e.g., DATIVE) used by various researchers to talk about their data and which refer to elements of the core ontology. Finally, there are the grounding concepts that are used to define the core linguistic concepts. The grounding concepts are really what exist in the world independent of language, including spatial concepts, temporal relations, abstract qualities, and physical objects. Figure 1 illustrates the organization of the major ontological components.

The grounding concepts are largely based on the Standard Upper Merged Ontology (SUMO), a specification of the IEEE Standard Upper Ontology working group (Niles and Pease 2001). Like other upper ontologies, the SUMO is meant to be a starting point for developing a specific domain ontology. We chose to use the SUMO as the base ontology for several reasons. It combines resources from many fields, for example James Allen's temporal logic (Allen and Hayes 1985), Beth Levin's verb class specification (Levin 1993), and John Sowa's upper ontology (Sowa 2000). The basis for describing linguistic objects is already present. It is freely available and extensible. It is open to review and critique by the knowledge engineering community.

Figure 1 illustrates just the backbone taxonomy (just *is-a* links) of the linguistic ontology. But there is much more in an ontology. Each concept is the ontology is defined according to a set of pre-defined predicates, a logic, and other concepts within the ontology. Take for example the concept of PastTense expressed by the label PAST as in (6).

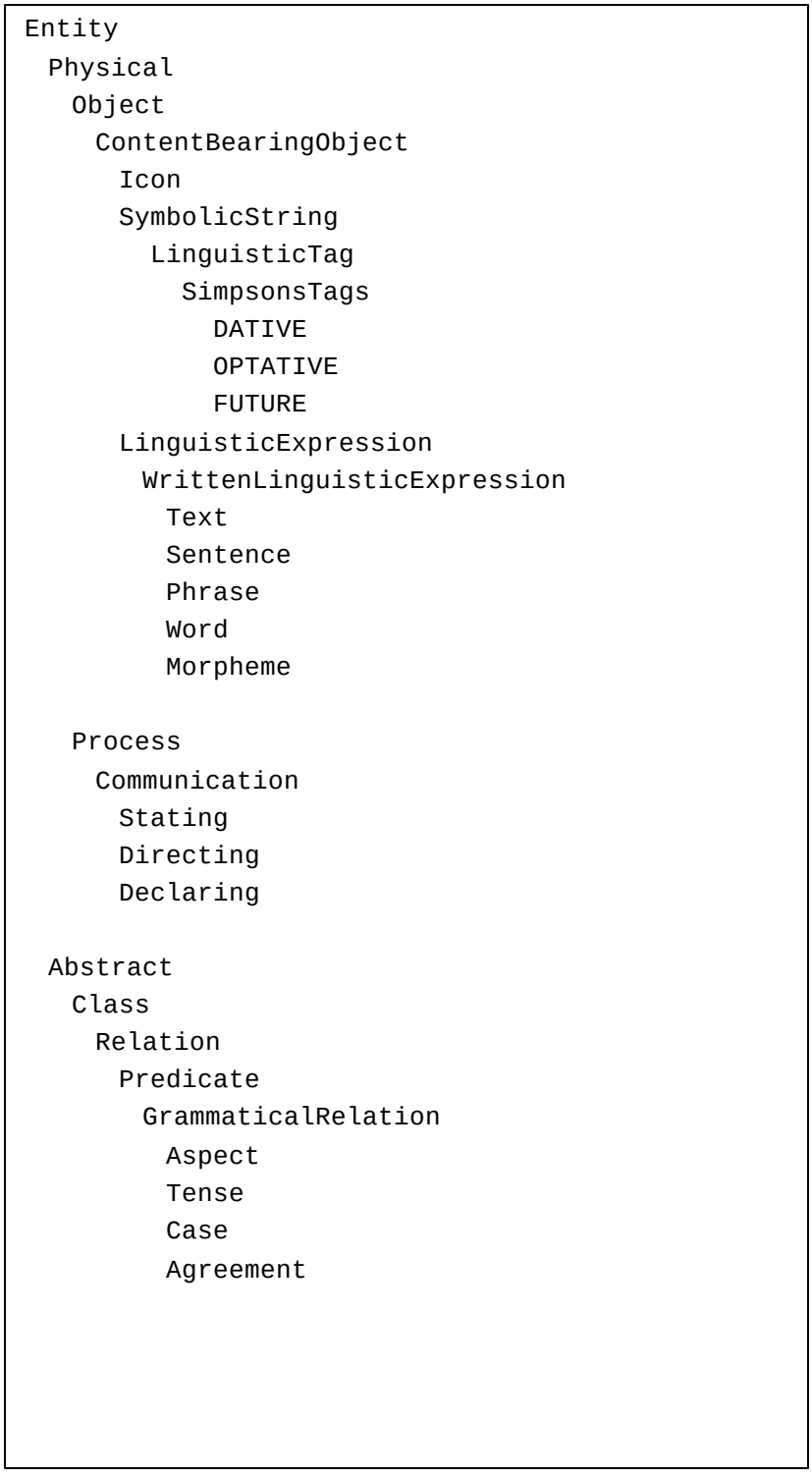

Figure 1. Major ontological components

(6) (<=>
```
        (past ?SENTENCE)
        (overlapsTemporally   (PastFn (WhenFn ?SENTENCE))
        (WhenFn ?EVENT)))
```
The logic statement in (6) means that some sentence is Past if and only if there is come process (event or state) described by the sentence that exists at a time before the time of speech. This is a very weak statement about past tense, but it may be derived from other more specific concepts, such as PastPerfective, PastContinuative, etc. Making very general statements like this also allows diverse data sets to be mapped to one another, which is the basis for interoperability. The more specific a statement the less likely it is to apply across data sets. Some concepts, such as Word or Morpheme may apply uniformly across all data sets, but, again, maybe not. In short, the ontology provides the necessary machinery to define *any* concept expressed in a linguistic analysis.

In order to develop the ontology, we first worked in a top-down manner identifying the major conceptual domains of linguistic analysis. Most importantly is the domain of the segment, giving rise to the concepts Discourse, Sentence, Phrase, Word, Morpheme. Then there are the major aspects of analysis including Case, Tense, Mood, and Aspect. The ontology was also constructed bottom-up. That is, we looked at the data in order to motivate ontological distinctions.

EMELD as part of the Semantic Web

In the terminology of the Semantic Web, the linguistic data community is a *resource* community. Such communities share data resources and tools and use similar vocabulary, or metadata, for describing their data. Metadata includes common attributes such as *author* and *language*, but also theory- and language-specific metadata such as morphosyntactic tags. As mentioned earlier, tagsets may and, in fact, do vary. What is needed is a mechanism for bringing resources together to achieve total data interoperability. The Resource Description Framework (RDF) is one such mechanism. RDF is becoming a central technology of the Semantic Web and is particularly well-suited for representing metadata about Web resources (Manola and Miller 2002).

RDF consists of resources and properties of those resources. RDF is often represented as a directed edge-labeled graph, as in Figure 2.

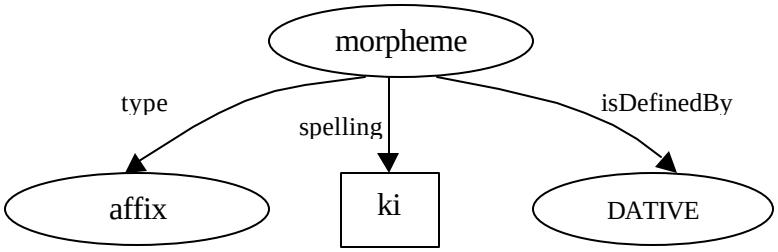

Figure 2. Directed edge-labeled graph representation of an RDF statement

The basis of RDF is the idea of a triple, such as in the statement "The DATIVE is defined by the linguistic ontology". What is really being expressed is the relationship between

some element (a tag) in a Web page and an element in the linguistic ontology (a concept). In RDF everything is either a resource or a property. A resource can be a Web page such as http://www.some-data.org/language_a.html/, some element of a Web page, such as DATIVE, or even a person, e.g., Australian language researcher Jane Simpson. Every resource is identified by a unique Uniform Resource Identifier (URI). This ensures that no resource will be confused with another. Properties describe some aspect of the resource. Properties too can be explicitly defined to avoid ambiguity. Properties are defined by using an RDF *Schema*. In the case of EMELD metadata, tags are linked via an RDF Schema to the ontology. In this way any search engine or other Web tool can be directed to a resource that specifies the semantics of the tag DATIVE. The RDF model can be rendered in various ways, including but not limited to labeled graphs and XML. Nodes and arcs and translated into machine-readable XML elements, attributes, and attribute values. An example is given in (7); the `<gram>` tag is replaced by the RDF isDefinedBy predicate.

(7) Boldfaced material in (5) translated into XML/RDF markup

```
<?xml version="1.0">
<rdf:RDF   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
           xmlns:onto="http://emeld.arizona.edu/ontology-ns#">
    <rdf:Description
           rdf:about="http://emeld.arizona.edu/morpheme"
           rdf:type rdf:resource="http://emeld.arizona.edu/affix">
       <onto:spelling>
           <rdf:Description onto:spelling="ki">
       </onto:spelling >
       <rdfs:isDefinedBy>
    <rdf:resource="http://www.emeld.arizona.edu/SimpsonsTags/DAT"/>
       </rdfs:isDefinedBy>
    </rdf:Description>
<rdf:RDF>
```

The real utility of using RDF is that it gives the researcher power to transform the data into a resource that is interconnected to many other resources on the Semantic Web. For our purposes, those resources include the ontology for defining linguistic concepts. But the power of RDF should not stop here. The field worker, for example, may wish to connect language data to pictures, sound recordings, or other cultural resources unique to the language community. In summary, RDF is a means of extending the data anywhere the Semantic Web can go.

The EMELD-Arizona system

The EMELD-Arizona system consists of an intelligent search engine with access to endangered language data on the Semantic Web. The language data is marked up in XML and conforms to the RDF data model enabling a precise semantics via the ontology. The search engine uses the ontology to reason about the language data. The search engine is a Web tool, while the language data and ontology are Web resources. The architecture is given in Figure 3.

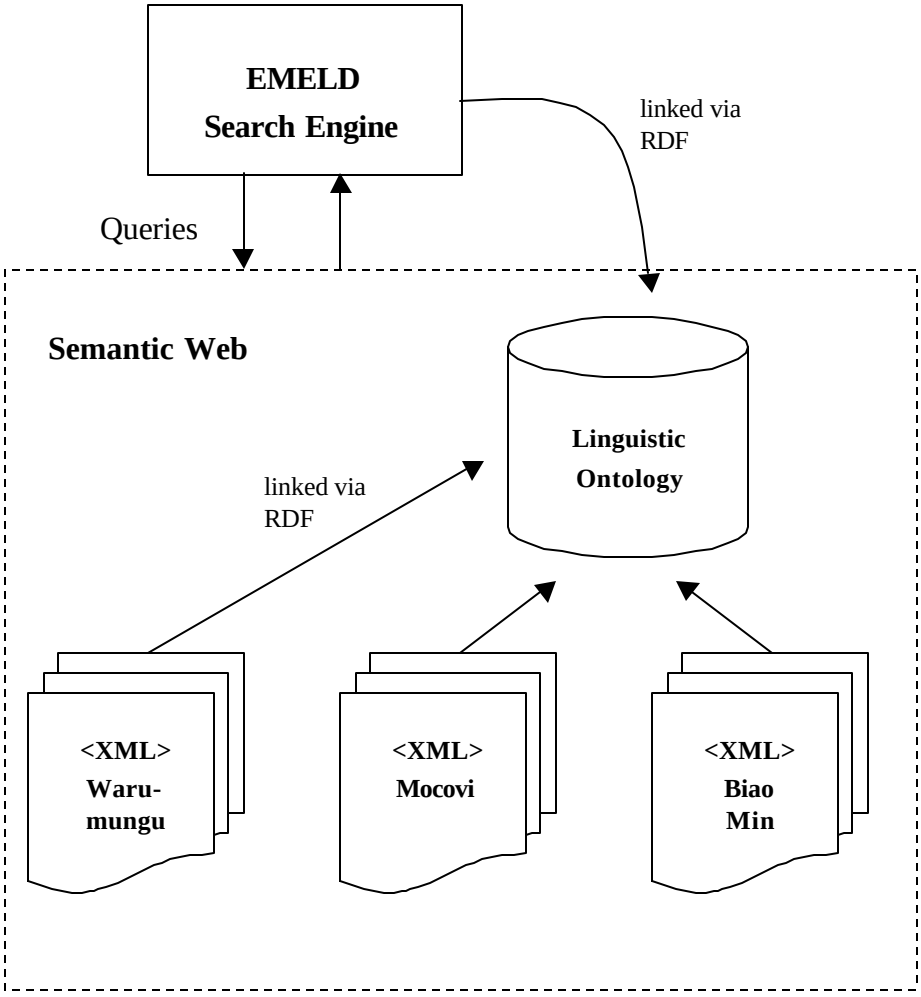

Figure 3: General architecture for the EMELD-Arizona system as part of the Semantic Web

The combination of these components results in an overall "smart" data model which translates into data interoperability and efficient searching. When developed, the EMELD-Arizona search engine can be used for a number of tasks. First, it can be used to return instances of particular markup, such as all uses of the label DAT. Second, it can return exemplar data based on some linguistic concept expressed in a number of ways over many data sets, for example, all instances of the perfective in Australian languages. Finally, it can be used to compare data across languages, like how the durative in Hopi differs from that of Mocovi. The first query type resembles the queries handled by traditional search engines available on the World Wide Web today. The second and third require the data model set forth for the Semantic Web. Comparing two languages in this way takes full advantage of the linguistic ontology and the reasoning power associated with the smart search engine.

Conclusion

We have presented an overview of the EMELD-Arizona system for smart search on the Semantic Web. The need for such a system is motivated by the variability of endangered language data. Central to the design of the search engine is the ontology of linguistic concepts. Together with its automated reasoning abilities, the search engine can perform a number of queries, including both traditional and "smart" queries. The EMELD-Arizona system is based on the data model for the Semantic Web and takes advantage of a number of Web technologies, including RDF and XML. As part of the semantic Web, the language researcher can place data in a fully extensible environment where information is given meaning and related to the ever growing body of electronic resources and tools. We argue that only with such a system can true data interoperability be achieved within the endangered language community.

References

Allen, James F. and Hayes, Patrick J. (1985) A common-sense theory of time. *Proceedings of AAAI-85*, 528-531.

Berners-Lee, Tim; Hendler, James and Lassila, Ora (2001) The Semantic Web. *Scientific American*, May 2001.

Lassila, Ora and Swick Ralph R. (1999) Resource Description Framework (RDF) Model and syntax specification. http://www.w3.org/TR/REC-rdf-syntax/

Levin, Beth (1993) *English Verb Classes and Alternations*. Chicago: University of Chicago Press.

Lewis, William; Farrar, Scott and Langendoen, D. Terence (2001) Building a knowledge base of morphosyntactic terminology. In *Proceedings of the IRCS Workshop on Linguistic Databases*, 150–156. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania

Manola, Frank and Miller, Eric (2002) RDF primer. http://www.w3.org/TR/rdf-primer/.

Niles, Ian, and Pease, Adam (2001) Toward a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (FOIS-2001).

Simpson, Jane (1998) Warumungu (Australian – Pama-Nyungan). In *The Handbook of Morphology*, ed. by Spencer, Andrew and Zwicky, Arnold M., 707-736. Oxford: Blackwell Publishers.

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