New Ways of Thinking about Lexical Resources: A Proposal for the Semantic Web

Scott Farrar
University of Arizona
February 9th, 2003

The following document was inspired by a working group held as part of the EMELD Workshop on Digitizing Lexical Information (2-5 August 2002, Ypsilanti, MI). It also reflects our recent thinking here in Arizona as we try to get a handle on the emerging Semantic Web.

1. Introduction

The current World Wide Web has the potential for becoming the primary means of storing and retrieving lexical data, if not linguistic data of all types. Publishing data on the Web allows the same lexical resources to be used for a variety of audiences: educators in indigenous communities, field linguists, theoreticians, software developers, and archivers. But, as is widely recognized, each of these groups requires a specialized model/schema so suit a particular purpose. Attempts to formulate a standardized model for the lexicon, such as the Text Encoding Initiative, have not been widely adopted because they do not provide a comprehensive structure that works for all of these audiences all of the time (see Ide, Kilgarriff and Romary 2000; Bell and Bird 2000). There are several extreme conclusions that lexicographers could draw from the present situation. First, a unified model is not feasible and its pursuit should be abandoned. Second, a unified model can be achieved from the merging competing ones resulting in a complicated model that allows for all types of variation. Third, a unified model can only be achieved if there is a fundamental shift in our thinking about lexical data.

As for the first conclusion we must concur, but only if approaches to the lexicon continue to be of the graphico-centric variety, that is, dominated by traditional print models, inspired primarily by dictionaries, less so thesauruses and encyclopedias (Bell and Bird 2000; Ide, Kilgarriff, and Romary 2002). For the lexicon, the natural unit of analysis is the lexical entry, or lexeme. On the other hand, there exists no ideal, agreed-upon lexical entry such that no all-encompassing lexicon model could ever be formulated. The second conclusion, then, must be rejected. Each part of the lexical entry (phonetic, phonological, orthographic, morphological, syntactic, semantic, pragmatic, etymological, etc.) is discrete, separable and recurring. This suggests a highly modular design with no privileged component—clearly a move away from the graphico-centric model. Bell and Bird (2000) recognize this to the extent of suggesting a data structure for each lexical entry, a 4-tuple <\text{Form}_i, \text{MSI}_i, \text{Sense}_i, \text{Aux}_j>, where each \text{Form}, \text{MSI} (morphosyntactic-info), \text{Sense}, and \text{Aux}(iliary information) can be separately identified and then combined depending on the intended use. However, the model proposed by Bell and Bird (2000) is too abstract, ultimately begging the question of how the elements of the 4-tuple will be mapped together to form a coherent whole. But if the lexicon is reconsidered with respect to the possibilities provided by the World Wide Web, especially the Semantic Web, then we argue that the mapping is made clear and a unified model can be achieved. Furthermore, we believe that our approach is applicable to all types of linguistic data, and thus argue for what we call the Unified Model for Language Data (UMLD).
The following lays out our vision for new ways of thinking about the lexicon. One of our goals is to show that the approach of Bird and Bell (2000) is adequate as a conceptual basis for the UMLD, but that a full implementation requires the framework of the Semantic Web. Through a few illustrative examples, we show how linguists can rapidly deploy complex lexical resources by building on simpler ones. We argue against an autonomous, standalone lexical entry and thus the entire graphico-centric paradigm. We claim that our approach encourages higher standards of practice than before, which will in turn help to ensure that lexical data is preserved. Furthermore, the desiderata for multimedia content as expressed in Gibbon (2000) can be realized as part of the Semantic Web.

2. Lexicon as Semantic Web Resource

As was mentioned at the beginning of this paper, the current World Wide Web has the potential to become the primary source for storing and accessing data. At the moment, however, the Web is designed to be human readable; its content is marked up in such a way that Web browsers can display data in a meaningful way to humans. As such the Web is not meant for machine consumption. For example, consider the difficulty that current Web search engines have in processing search queries. While search engines may return impressive results, their levels of precision and recall fall far short of the standards that are expected of information extraction systems that operate over much more structured data. The limitations of search capability preclude the use of natural language interfaces or structured query formats. Furthermore, searches over data that are anything but text-based are simply not possible. The preceding sentences characterize first generation of Web technology. Already, though, we are moving toward a second generation Web model—the Semantic Web (Berners-Lee 1999; Berners-Lee, Hendler, and Lassila 2001).

The requirements for the Semantic Web include 1) new technology in markup languages, 2) a Web infrastructure formed around meaning (not display), and 3) new tools for navigating, searching, and publishing Web data (ibid). In this section, we focus on the second requirement for the Semantic Web, using the markup of linguistic data as a test case. The required Web infrastructure includes semantic resources that will act as repositories of knowledge about particular domains. The primary semantic resources on the Web will be ontologies, enabling the Web to be transformed into a highly structured knowledge system: “[T]he… Semantic Web hinges on the possibility to use shared vocabularies for describing resource content and capabilities, whose semantics is described in a (reasonably) unambiguous and machine-processable form” (Masolo et al. 2002: 6). A Web ontology is intended to be much more than just a controlled vocabulary, that is, more than just an agreed-upon set of values for elements in a markup language. Ontologies on the Semantic Web will require “explicit representation of ontological commitment” and “rich axiomatization” (ibid). According to Masolo et al. (2002), the ontologies envisioned for the Semantic Web fall into two categories: lightweight ontologies (often just taxonomies) that provide “semantic access to a specific resource” in which the meanings of terms are already known; and foundational ontologies that are to be used to help Web agents “negotiate meaning” and to establish a consensus between human users and agents. As part of the EMELD project, the group at Arizona is involved in the creation of one such foundational ontology, the General Ontology for Linguistic Description (GOLD) (Farrar and Langendoen, to appear). We believe GOLD will act as a pivotal resource for creating
communities of practice, or groups of Web users with common goals and resources. Besides ontologies, semantic resources include terminology sets (familiar terms used by a particular group, e.g., Bantuists, Quechua linguists), folk taxonomies (language-specific worldviews), and compilations of theory-specific constructs (e.g., semantic primitives). The idea is that GOLD, and perhaps lightweight ontologies, will provide the metastructure (see Simons 2002) necessary for establishing the linguistics “corner” of the Semantic Web. In practice such resources will be the catalyst for the rapid deployment of lexical resources and foster high-quality markup practice.

The following presents a scenario in which an existing lexicon is used to construct one of even more complexity. Consider the case of the Huallaga Quechua Dictionary (Weber 2002) that incorporates the folk taxonomy in Figure 1:

```
Top 'everything'
  Cawajcuna 'what lives'
  …
  Animal 'animal'
    Abicuna 'birds'
  …
  Chucaru 'wild'
Mana cawajcuna 'what is not alive'
  …
  Mana charina caj 'what cannot be grasped'
  …
  Yacu 'precipitation'
```

**Figure 1—Huallaga folk taxonomy, based on culturally received worldview**

Lexical entries in the Huallaga print dictionary, namely the gloss material, include reference to elements of the taxonomy. In a Web version, the taxonomy could be a stand-alone resource that is only referenced in the main lexicon. The Huallaga taxonomy is not unlike the basis on which WordNet is organized. As WordNet’s categories are accepted at least within a specific community, the Huallaga taxonomy could serve as a Quechua community resource. And just as WordNet has facilitated the development of NLP tools, the Huallaga taxonomy could be used for subsequent lexicographic efforts. What is needed for the scenario to work is a markup standard that allows the modular insertion of references to the taxonomy and, as will be adduced, other linguistic resources.

Going a step further, what if everything including \(<Form_i, MSI_j, Sense_k, Aux_l>\) were modular? In practice, then, a lexical entry would be “virtual” in the sense that it would be a set of pointers to various document-internal and document-external resources. The model that emerges can be schematically represented in Figure 2.
Figure 2—The lexical entry as virtual Web resource

From the figure, the lexical entry consists of different types of information, with no particular kind of information privileged over another. Also, this model requires that the hyperlinks, represented in the figure with arrows, have different meanings. The arrow to the taxonomy box should represent the ‘is-a-kind-of’ relation, while the arrow to acoustic info should represent the ‘is-pronounced’ relation. No benefit would be gained from simply pointing to the various external resources; HTTP protocol provides only the linking mechanism with no means of specifying a semantics. Resource Description Framework (RDF) is the markup standard that gives a link a precise semantics. A resource is anything that can be identified on the Web, such as a webpage, sound file, or an individual XML element/value. RDF provides the mechanism to make statements about Web resources—an information model meant to convey that something represents something else. Consider the example in (3).

(3)

```xml
  line 1   <gold:Word rdf:ID="lex002">
  line 2      <rdfs:label xml:lang="x-sil-HGA">cutichi</rdfs:label>
  line 3      <gold:gloss xml:lang="en">to return something...</gold:gloss>
  line 4      <gold:derivedFrom rdfs:resource="lex001"/>
  line 5      <gold:pronounced rdfs:resource="#HallaguaSound"/>
  line 6   </gold:Word>
```

As the RDF example is meant to be machine readable, it deserves some deal of explanation. First off, RDF is a structuring format (a syntax) for markup and is itself well-formed XML. Therefore it relies extensively on the use of namespaces. For example, many of the elements are prefixed with ‘gold’. This simply indicates that the associated element is part of another XML document, in this case the GOLD ontology mentioned above. RDF enables a standard way of making statements about Web resources, in the form of subject—predicate—object. The subject in this
case is identified on line 1 as the WORD with the unique id "lex002". What follows in lines 2-5 are various predicates-object pairs: the entry is labeled as cutichi; the entry is glossed in English as “to return something…”; the entry is derived from the entry lex001; the entry is pronounced as [sound file in HuallagaSound]. What is important to notice about RDF’s information model is that a Web resource (a subject or object) is allowed to be anything that can be symbolically referred to on the Web—a recording, an event, or an idea—not just webpages, strings or numbers. This is why RDF is so useful as a model for metadata, although as (3) illustrates, RDF can serve as much more.

The development of XML and RDF 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. An emerging Web standard for encoding foundational ontologies is the Web Ontology Language, referred to as “OWL”\(^2\). Web ontologies are ontologies in the artificial-intelligence sense in that they provide the means of making explicit a conceptualization of a domain, declaring what entities and relations exist. Example (4) shows the use of OWL in defining the class PREFIX.

(4)

```xml
<owl:Class rdf:ID="Prefix">
  <rdfs:subClassOf rdf:resource="#Affix"/>
  <rdfs:label xml:lang="en">Prefix</rdfs:label>
  <rdfs:label xml:lang="fr">Préfixe</rdfs:label>
  ...
</owl:Class>
```

The statement declares PREFIX to be an OWL class and a subclass of AFFIX. It also declares PREFIX to have the English label “Prefix” and the French label “Préfixe”. Again, OWL is of the format subject—predicate—object, just like RDF. A class, such as the one given above, would also contain rules of inference that would be useful to an expert system that reasoned about linguistic data. We are currently enriching GOLD with categories relevant to lexical resources.

Therefore the requirements for the Unified Model of Linguistic Data begin to emerge: for semantics, ontologies, folk taxonomies, and terminology sets are crucial. But what about other types of information, the Form, MSI, and Aux in the model of Bell and Bird (2000)? We have glossed over some of these details of the model. In order for the UMLD to work, we assume that the following standards will be in place: orthographic encoding standards (e.g., Unicode), standards for multimedia encoding (see Gibbon 2000), standards for metadata (cf. the controlled vocabularies of IMDI and OLAC), and finally standards for morphosyntactic information in the form of feature systems (cf. Langendoen and Simons 1995; Ide and Romary 2002).

3 Some Examples

In the best of all worlds, a lexicon should contain as much information as possible. Transformation technologies (such as XSLT) can then be used for rendering down when a simple typographic display is needed (cf. Ide, Kilgarriff and Romary 2000). Taking to the

\(^2\) See http://www.w3.org/2001/sw/
extreme the proposal of Bell and Bird (2000), we argue that the lexical entry is merely a construct of presentation. If certain components of $<Form_i, MSI_j, Sense_k, Aux_l>$ are grouped together in the output of a transformation, then result alone is the lexical entry. In some cases, it may include only form and translation, in others only pronunciation and meaning. What this allows is the freedom to structure entries in a custom manner while keeping the abstract underlying structure. Consider, for example, the difficulty of reconciling lexicons that have totally different hierarchical models. Referring to Bell and Bird (2002), in an Orokolo dictionary, reproduced here as (4), “the same orthographic form and part of speech are represented as three independent entries”, while one entry contains a sub-entry based on a de-nominal.

(4)

\begin{itemize}
  \item para\textsuperscript{1}\ldots n. a kind of mangrove.
  \item para\textsuperscript{2}\ldots n. club.
  \item para\textsuperscript{3}\ldots n. an oar.
  \item para koa\ldots v.t. to ply oar, to row.
\end{itemize}

This can be compared to a Sango dictionary “which combines multiple sense definitions into a single entry” reproduced in (5).

(5)

\begin{itemize}
  \item ngôrô\ldots
    \begin{enumerate}
      \item entourer, encercler, contourner\ldots
      \item accueillir un visiteur, aller au devant d’un hôte\ldots
    \end{enumerate}
\end{itemize}

In the UMLD, a possible underlying structure for the Orokolo example is given in (6).

(6)

\begin{verbatim}
line 1  <gold:Word rdf:ID="lex002">
  line 2  <rdfs:label xml:lang="x-sil-ORK">para</rdfs:label>
  line 3  <OrkoloTermSet:POS>noun</OrkoloTermSet:POS> a kind of mangrove</gold:gloss>
  line 4  <gold:gloss xml:lang="en">oar</gold:gloss>
  line 5  <gold:pronounced rdfs:resource="#OrkoloTranscription"/>
</gold:Word>

line 9  <gold:Word rdf:ID="lex003">
  line 10 <rdfs:label xml:lang="x-sil-ORK">para</rdfs:label>
  line 11  <OrkoloTermSet:POS>verb</OrkoloTermSet:POS> to ply oar</gold:gloss>
  line 12  <gold:gloss xml:lang="en">to ply oar</gold:gloss>
  line 13  <gold:derivedFrom rdfs:resource="#lex002/gloss[3]"/>
</gold:Word>
\end{verbatim}

With the underlying model, the Orokolo data can be structured into a lexical entry by an (XSLT) transformation engine either as separate entries as shown in (4) or combined as in the Sango case, in (5). Now notice that even though we have made a distinction based on part of speech, the nouns are given in lines 1-8 and the verbs in 9-14, the separation could have been based on
sense, rather than part of speech. Either way, the final rendering via XSLT determines the make up of the lexical entry. The hierarchical relation between the particular noun sense and the verb can be instantiated during the transformation using XPath (see the XPath expression in line 13). Furthermore, even though we have chosen to separate lex002 from lex003, RDF is sufficiently flexible as to allow them to be combined. By our separating them, it should be inferred that there are two separate lexical entries. The resources lex002 and lex003 merely represent abstractions.

Turning to another kind of example, Bell and Bird (2000) note the difficulty that their model has with hypothetical entries such as the English one in (7):

(7)

establish: to set up…
   establishment: something which has been set up…
       disestablishment:…
           disestablishmentarian:…
               disestablishmentarianism:…
                   antidisestablishmentarianism:…

Whereas this kind of entry seems “perverse” (according to Bell and Bird 2000), such kind of hierarchical, recursive structure is a natural device in agglutinating languages, as demonstrated by Weber (2002). Consider the Huallaga entry in (8).

(8)

 cuti- v.i. ‘to return’
   cutichi- v.tr. ‘return (something to someone)’
       ayńita cutichi-
           1. v.tr. ‘help another in exchange for help given’
           2. v.tr. ‘to take revenge, to get even with’
   cutipa
       1. s.n. ‘relapse’
       2. s.n. ‘second cultivation of corn…’
   cutipa- v.i. ‘to repeat (especially referring to cultivation…’
   cutipä- v.tr. ‘to return to a woman after having left her’
   ...

Weber (2000) argues convincingly that such entries should be accommodated in any purportedly standard model for lexical encoding.

(9)

 line 1  <gold:Root rdf:ID="lex001">
 line 2     <rdfs:label xml:lang="x-sil-HGA">cuti</rdfs:label>
 line 3     <gold:gloss xml:lang="en">to return</gold:gloss>
 line 4  </gold:Root>
Notice that there is a difference in the subjects of the various RDF statements: one is ROOT, while others are STEMS or WORDS. Critically, the examples in (9) can be sorted by part of speech and derivational hierarchy to achieve the rendering in (8) as long as the appropriate RDF statements are included. It is beyond the scope of this paper to give examples where rendering is based on the sorting of morphosyntactic or phonological features. However, we see this as no more difficult than the examples in (5-8). We also recognize that a thorough treatment of these examples should include a discussion of actual XSLT scripts.

4 Conclusions

To summarize, we argue that the traditional notion of the autonomous lexical entry in the context of a graphico-centric paradigm should be abandoned. We embrace the proposal of Bell and Bird (2000) as an abstract data model for lexical encoding and for the encoding of the full range of linguistic data. By pushing the model further and by incorporating the technology of the Semantic Web, we propose a Unified Model for Linguistic Data. We have argued that the Web has the potential to become the primary means of storing and accessing linguistic data. We show by example how the UMLD will be instantiated as part of the Semantic Web. We argue that this model will facilitate the rapid development of lexical resources.
**Works Cited**


