An Ontology for Linguistic Annotation

Scott Farrar, William D. Lewis, and D. Terence Langendoen
Department of Linguistics, University of Arizona
{farrar, wlewis, langendt}@u.arizona.edu

This paper discusses some of the design criteria for a linguistic ontology that can be used to support multilingual and crosslinguistic searches and queries on the Internet. It focuses on integrating linguistic concepts and instances into an upper-level ontology, and shows that the result can be understood and analyzed as a feature (structure) system. It considers various types of linguistic structure ranging from segment types to grammatical properties and relations, and linguistic inventories including phoneme tables, inflectional paradigms, lexicons, and grammatical descriptions.

Introduction

This paper addresses two of the five topics in the call for proposals for this workshop:

- Problems for representing linguistic data;
- Representing meaning in natural languages using ontological support.

We contend that these are really two aspects of the same topic. The notion "linguistic data" can and should be construed broadly enough to encompass the notion of "meaning", along with every other aspect of the structure of a language that is of linguistic interest, including its phonetics, phonology, lexicon, morphology, and syntax. Ontological support is needed for all kinds of linguistic data on the internet, not just for semantics.

We approach the problem from the perspective of comparing data from and about many different languages. Websites containing documents of various sorts are being created for many languages from around the world, some by the communities that speak these languages and others by linguists who study them. These documents include teaching materials, word lists, dictionaries, plain texts (some with translations into another language), glossed texts, grammatical descriptions, and audio and video clips (some annotated). In order to be able to use these documents for linguistic analysis or for information retrieval, a means for comparing them must be available.

The core of any application for comparing linguistic data is an ontology of the elements of linguistic analysis at every level from phonetics to semantics, including the states and events of common human experience and the components of those states and events (participants, things, actions, properties, locations, etc.). As such, the ontology must include knowledge of both linguistics and specific languages. In our part of the NSF-supported Electronic Metastructure for Endangered Language Data (EMELD) project, we are in the process of developing such an ontology <emeld.douglass.arizona.edu>. We are building on the Standard Upper Merged Ontology (SUMO) of the IEEE SUO working group <suo.ieee.org>, which is designed to be extended to encompass domain-specific ontologies. The following section discusses some of the extensions and modifications to SUMO that we are making to create our ontology. In this endeavor, we have made every effort to follow the "OntoClean" methodology of Guarino & Welty (2002). In addition, we show the relation between the structure of ontologies and that of feature systems.

Toward a Linguistic Ontology

The first step in constructing an linguistic ontology is to establish backbone taxonomies for the domain. The major subdivisions within the ontology include linguistic segments, grammatical properties and relations, and inventories. We consider each of these taxonomies in the following subsections.

Linguistic Segments

SUMO already contains basic segmental notions such as Word, Phrase and Sentence subsumed under the concept LinguisticExpression. However, we are making significant revisions to this section; some of our preliminary results are shown in Figure 1.

```
LinguisticExpression
   Morpheme
      BoundMorpheme
         Root
         Affix
            Prefix
            Suffix
      FreeMorpheme
   WordPart
      Stem
   Word
      FunctionWord
      ContentWord
   Phrase
   Sentence
   Text
```

Figure 1 Partial taxonomy of linguistic segments

The classes in Figure 1 are adequate to describe most segmental phenomena encountered in the domain of morphosyntax, which is where we are starting our work. For example consider the snippet of Warumungu glossed text in Figure 2 (from Simpson 1998: 727, example 19). From the conventions of text glossing (which can be represented in various ways in a markup language) and rudimentary knowledge of Warumungu grammar, we can infer that *purrumu* is an instance of Root and *rra* of Suffix (and from the ontology, also of BoundMorpheme, Morpheme, and LinguisticExpression). In addition, *purrumurra* as a whole is an instance of ContentWord (and hence also of Word) and of Sentence.

```
Purrumu-rra!
touch-IMPER
‘Touch it!’
```

Figure 2 Simple Warumungu glossed text

Of course there is much more to the analysis of these segments than identifying their segment type. For example they are segments in the Warumungu language, not in Sanskrit or C#. We can use an ontology to express these facts as well. Language is a category in SUMO, and Warumungu can be considered an instance of that category. Moreover, the strings *purrumu* and *rra* are written in an orthography designed for that language, and so can be considered instances of the category OrthographicString (or OString). This category is a kind of SUMO Relation, since it has two arguments, the first a string of characters and the second the orthography in which that string is written; it should also be considered a kind of SymbolicString, a category already defined in SUMO. We propose to add Orthography to the ontology defined as a set of characters with a sort order defined over them, and other properties. WarumunguOrthography (or WOrth) is an instance of Orthography. Consequently the instance of Root in Figure 2 is the structured object representing the combination, by mutual inheritance, of Language and OString concepts and instances in Figure 3.

```
      +-                          -+
[Root | Language Warumungu         | ]
      | OString  <purrumu WOrth> |
      +-                          -+
```

Figure 3 Warumungu *purrumu* as a structured instance of Root

The relation between ontological categories and instances is essentially that of feature (structure) systems (Shieber 1986; Langendoen & Simons 1995; Maxwell, Simons & Hayashi 2002). A category and its instance together is a feature, with the category being a feature name and the instance a feature value. An instance, or feature value, may be unstructured, such as that of the category, or feature names, Language; or a feature structure, such as that of Root. Unstructured instances may be further typed, such as

boolean, integer, string, member of a closed class of atoms (as in the case of instances of Language), etc. The value may also be a list of instances, as in the case of OString. Those instances may themselves be unstructured or feature structures.

Finally, we note that a property in SUMO (one-place Predicate, called Attribute) is represented as having a single instance, and Relation in SUMO (a multi-place Predicate) as having a list of two or more instances.

The Root feature in Figure 3 could in principle be made part of the linguistic ontology, just like any other instantiation of an ontological category. However, the most appropriate place for its value to be stored is as a lemma in a machine-readable dictionary of Warumungu, together with a link to the ontological category of which it is a value, and links to its occurrences in Warumungu texts and other documents.

The analysis of *purrumurra* as a Word must take into account that it is composed of the parts *purrumu* and *rra*, and so should contain the features that represent those parts, as in Figure 4 (with additional abbreviations to make the figure fit the allotted space).

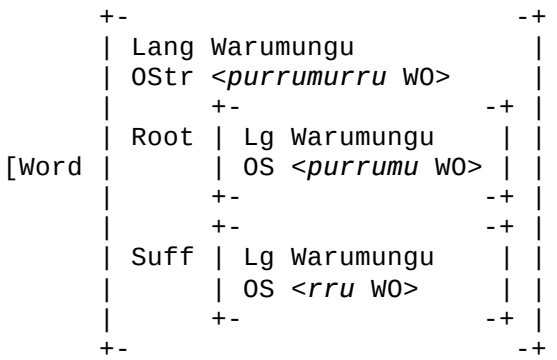

Figure 4 Warumungu *purrumurru* as a structured instance of Word

The structure in Figure 4 appears redundant, since (in particular) the Language feature appears in three different feature structure components. The "percolation" of this feature can be expressed by a rule that in effect specifies that the Language feature of a Word is the unification of the Language feature of its parts. The spelling of the word is also predictable from the spelling of its parts together with the fact that *rru* is a Suffix.[1]

Grammatical Properties and Relations

We have already observed that SUMO distinguishes between properties, subsumed under the category Attribute, and relations, subsumed under Relation. Among the properties already identified in SUMO are ColorProperty and ShapeProperty, to which we have added GrammaticalProperty. Among the relations already identified in SUMO are SpatialRelation and TemporalRelation, to which we have added GrammaticalRelation.

From the information in Figure 2 and knowledge of Warumungu grammar, we conclude that *purrumu* is a verb root that expresses the two-place Relation 'touch'; we leave open the question of what subclass of Relation it is an instance of. The fact that a Root that means 'touch' is a verb root is not an accident; we would not expect a verb root in any language to mean 'kangaroo'. To account for this, we propose first that Express is a GrammaticalRelation that holds between a LinguisticExpression and its meaning (in this case another two-place Relation). Second, we define Meaning as a GrammaticalProperty of any LinguisticExpression that is a first argument of Express; the value of the Meaning feature being the second argument in a true assertion of that Relation. Third, we propose that GrammaticalCategory (henceforth Category) is another

[1] Suffix is represented in Figure 4 as an Attribute, but should really be considered a Relation, whose second argument is the LinguisticExpression to which it is suffixed. In this case, it is the Root *purrumu*, which can be indicated by a "reentrant value", a pointer to the occurrence of that root in the Word feature.

GrammaticalProperty, and that one of its instances is Verb. Finally, we propose that any LinguisticExpression can have a Category feature, but that there are rules that relate its other features, including what it expresses, to the value of that feature.

We also conclude that *rra* is a Suffix that expresses the three-place Relation 'command', in which the first argument is the Speaker, the second argument is the Hearer, and the third argument is a Proposition about (in this case) the Predicate (Attribute or Relation) expressed by the verb root to which it attaches. This Predicate is further understood as holding of the Hearer. It is conventional to associate the feature `[Mood Imperative]` with such expressions, so that the presence of this feature is a surrogate for the Meaning of the expression containing it.[2]

Mood is a subclass of GrammaticalProperty whose other instances include Declarative and Interrogative, and so represents a range of GrammaticalRelations involving the Speaker, the Hearer, and the Proposition being expressed. The Mood feature also "percolates" to the Word and Sentence containing the Morpheme with that feature, with its argument variables instantiated. The grammatical properties of the linguistic expressions in Figure 2 appear in the SUMO Attribute hierarchy fragment in Figure 5.

In general, a LinguisticExpression can be viewed as a feature whose value is a feature structure consisting at least of Language and SymbolicString features, and some number of GrammaticalProperty features as in Figure 6, which extends the analysis in Figure 3. For practical reasons described below, we propose that grammatical relations appear only indirectly in such feature structures via their GrammaticalProperty counterparts, for example the Express Relation appears via the Meaning Attribute.

```
Attribute
   GrammaticalProperty
      Meaning
      Category
         Noun
         Verb
      Mood
         Declarative
         Imperative
         Interrogative
```

Figure 5 Place of Category and Mood in the SUMO Attribute hierarchy

```
      +-                          -+
      | Language Warumungu         |
[Root | OString  <purrumu WOrth> | ]
      | Meaning  'touch'           |
      | Category  Verb             |
      +-                          -+
```

Figure 6 Further analysis of Warumungu *purrumu* as a structured instance of Root

Our proposal that grammatical relations be represented indirectly as grammatical properties is meant to apply only to the representation of expressions in glossed texts, lexicons and the like, not for feature-structure-based linguistic analyses generally, particularly those designed to parse or generate text, rather than simply to represent it. We have no quarrel, for example, with HPSG analyses in which the arguments of a multiplace predicate are represented as a list, so that the relational character of that predicate is explicitly represented (Pollard & Sag 1994). Our recommendation is simply that in the web-based linguistic ontology, a means be provided for coding grammatical relations as attributes of the linguistic expressions that enter into those relations. So for example, the Warumungu sentence *Purrumurra!* 'Touch it!' would have a Subject attribute (a feature) whose value is a feature-structure representing the Hearer, and a Direct-object attribute

[2] The label IMPER that Simpson uses for this feature value is significant only for the document itself. The ontological category should be linked to the term that the label stands for, not the label itself.

whose value is a feature structure representing an Object, whose reference may be determined in the larger text in which that sentence appears.

Similarly, we propose that the three-place AssignCase relation, involving a CaseAssigner (typically a verb or function word of some sort, e.g. a preposition), a Case Assignee (typically a noun phrase bearing some GrammaticalRelation to the verb or function word), and a Case, be indirectly represented by a set of GrammaticalProperty features on the linguistic expressions for the CaseAssigner, the CaseAssignee and the Case. In Warumungu, the verb root *ngu* 'lie' assigns DativeCase to the NounPhrase that serves as its Direct-object, as illustrated by the example in Figure 7 (Simpson 1998: 727, example 22).

```
Yama+ajurnu  jarti-ki+karn
leave+3PL.NS other-DAT+now
pikapikka-ka ngu-nngara.
children-DAT lie-OPT.FUT
'Some should be left for the
other children.'
```

Figure 7 Illustration of DativeCase assignment in Warumungu

In this example, the suffixes *ki* and *ka* contain the feature `[Case Dative]`, which is inherited by the words *jartikikarn* and *pikapikkaka* containing them, and then by the NounPhrase *jartikikarn pikapikkaka.* That noun phrase in turn is represented as the Direct-object of the verb root *ngu*. From this representation, the fact that *ngu* assigns DativeCase to its Direct-object does not follow, but that fact could be represented explicitly as a feature of *ngu*, say as `[CaseOfDirectObject Dative]`. By omitting certain features that contribute to the AssignCase relation, such as CaseOfDirectObject, the individuals encoding the data do not have to commit to any particular theory of case relations in marking up their data for Case. Such markup could be used either by those individuals or other researchers to determine what theory of case relations best accounts for those data.[3]

Finally, the association of particular GrammaticalProperty features with particular LinguisticExpression types may be determined by general or language-specific rules. In Warumungu, for example, the feature `[Category Verb]` is directly associated with the Root type, whereas the feature `[Category Noun]` is directly associated with the Word type.

Linguistic inventories

Linguists present much of their analysis of a language in the form of tables and structured lists of various sorts, for which we use the cover term 'inventory'. These provide a handy form for expressing information about a language in a compact form, and with proper encoding provide a convenient source for capturing a great deal of linguistic data. In some cases, they also reveal conflicts in the use of terms that must be resolved if consistency and data integrity are to be achieved.

We propose that several of the types of inventories that linguists use be instantiated as concepts in the linguistic ontology, among them phoneme tables, inflectional paradigms, and lexicons. By doing so, we hope to achieve a degree of uniformity and consistency that is currently lacking in the linguistics literature, but without imposing an encoding standard that will simply be ignored by the majority of data providers.

[3] In the suggested encoding, the two-place grammatical relation AgreeInCase that presumably holds between *jartiki* and *pikapikkaka* is also not represented.

Phoneme Tables

A very familiar datatype is the consonant phoneme table, with the rows representing manner of articulation (typically from least to most sonorant), and the columns points of articulation (from lips to glottis). The phoneme that appears in a particular cell has the manner-of-articulation feature(s) of its row and the point-of-articulation feature(s) of its column. So for example from the fragment of the Warumungu consonant phoneme table in Figure 8, simplified from Simpson (1998: 710, Table 32.1), we learn that the WOrth segment *pp* represents two distinct phonemes, one a long voiceless bilabial stop [p:], and the other a short one [p].

	bilabial	*alveolar*	*...*
stop			
long vl	pp [p:]	tt [t:]	
short vl	pp [p]	tt [t:]	
short vd	p [b]	t [d]	
nasal	m	n	
lateral		l	
flap		rr [r]	
semivowel	w		

Figure 8 Fragment of Warumungu phoneme table

Simpson's table is an instance of the ontological subcategory ConsonantPhonemeTable, defined as providing manner- and place-of-articulation information about the consonant phonemes of a language in the manner just described. The features that appear in the row and column identifiers in any instantiation of the table should also be linked to the linguistic ontology.

Inflectional Paradigms

Perhaps even more familiar are inflectional paradigm tables such as conjugations for person and number of the subject in various tenses and aspects of the Spanish verb as in Kendris (1990), and declensions for case and number in the various Latin noun classes. As these examples illustrate, the category InflectionalParadigm is not defined by its features, since these vary from paradigm to paradigm. Rather, it is characterized by:

1. *completeness*: every instance of the combination of interacting features is instantiated;
2. *multidimensionality*: at least two features are involved;
3. *significance*: the features play a major role in the grammar of the language;
4. *contrast*: the linguistic expressions that manifest the feature combinations (typically morphemes) contrast with each other, typically in a given position within a word.

The completeness criterion gives rise to the concept of Suppletion, the occurrence of phonologically distinct forms in parts of the paradigm, for example *voy* 'I go', *fui* 'I went' and *iba* 'I was going' are all part of the same paradigm of the Spanish verb *ir* 'to go'. It also accounts for the description of paradigmatic gaps as unexpected, as the absence of an 'away' form of the L-conjugation class of verb roots in the past continuative tense (sic) in Warumungu (Simpson 1998: 722, Table 32.4). Similarly, incomplete paradigms are sometimes described as "defective".

The contrast criterion sometimes appears to make strange bedfellows. For example, in Warumungu, the following instances contrast in the paradigm for motional verbs: future, imperative, present, past punctual, past continuative, admonitive, and optative/irrealis (Simpson 1998: 723, Table 32.5). These appear to be instances of four different features: Tense (past, present, future), Aspect (punctual, continuative), Mode (optative/irrealis) and Mood (imperative and admonitive). The solution in this case is to identify the nearest common feature in the ontology for Tense, Aspect, Mode and Mood, so that the contrasting instances are all instances of that common feature.

Lexicons

At minimum, a lexicon consists of a list of expressions in a given language, and some grammatical features associated with each, usually including its meaning. In a monolingual lexicon the meanings are expressed as paraphrases in the same language; in a bilingual lexicon as a translation into another language. Given an ontology of concepts that provide a basis for the analysis of meaning of expressions in human languages (see Gangemi et al. 2001), we can envision lexicons in which each meaning is an instance of a universally defined feature or bundle of features, i.e. a feature structure, that can be compared systematically with meanings of other entries in the same lexicon, or in different lexicons, i.e. of different languages.

Moreover if indeed a grammar is simply a means of projecting a lexicon onto a language as a whole (i.e. all of its expressions), then access even to incomplete (but well designed) lexicons for a variety of languages on the Internet that are correctly linked to the linguistic ontology we envision will be a tremendous resource for both language comparison and multilingual information retrieval.

Two Goals for a Linguistic Ontology

Throughout this paper, we have made more or less explicit certain concepts that may be described as "grammatical" in the narrow sense of that term, whose understanding requires knowledge of delimited, but still extensive, aspects of semantics as a whole. We have however alluded to the possibility that a linguistic ontology that deals with all of linguistic semantics, including lexical meaning in its full glory, can be constructed. We now consider two research goals, which can be worked on in parallel, though not entirely independently of each other.

The first goal is to define each narrow-sense grammatical concept and its possible instances, such as DurativeAspect and InstrumentalCase; in effect to provide the semantics of the closed-class vocabulary of as many languages as we can get adequate descriptions of. The second is to provide the basis of the analysis of the open-class vocabulary of human languages along the lines sketched out toward the end of the preceding section. The reason that these goals are not independent is that the distinction between features for open and closed classes is not a clear one in any one particular language, and even less so when comparing different languages. For example, the features involved in the analysis of the closed class of "handling verbs" in Athapascan languages are very much like those needed to describe open-class vocabulary items in other languages.

SUMO has certain properties that may help us toward reaching the first goal. We may suppose that the semantics of Tense and Aspect features of human languages can be formulated in terms of the attributes and categories of temporal logic. Since SUMO already implements the temporal logic of Allen & Hayes (1985), including such notions as Before, During, TimeInterval, and TimePoint, we can formulate within SUMO some general definitions concerning Tense instances, as in Figure 9 in SUO-KIF format.

```
(<=>
  (present ?SENTENCE)
    (and
      (exists ?PROCESS Process)
      (represents ?SENTENCE ?PROCESS)
      (overlapsTemporally
          (WhenFn ?SENTENCE)
          (WhenFn ?PROCESS)))
```

Figure 9 A definition for PresentTense

This definition may be read as "A sentence is present tense just in case there exists some process such that the sentence represents that

process and some part of the process occurs at the time of speech."

If the terms and the logic of the ontology are rich enough, the very subtle distinctions that are found in the closed-class semantics of human languages can be adequately expressed, as long as the descriptions we have at hand are clear enough for us to understand what they are. Suppose in the description of two different languages, we find a bound morpheme in each labeled "durative", described as applying to a state or activity that lasts for a relatively long time interval. However upon closer examination, we discover that they actually mean different things: One requires that the state or activity be continuous over the interval, and the other does not. Clearly these are distinct instances of the Aspect category.

Reaching the second goal requires constructing a very broad-coverage ontology that is designed specifically to work with the full range of human language data. The ontology designed for the CYC Project (Lenat & Guha 1990) may be adaptable for this purpose. We have chosen to work with SUMO, because we think we understand it better, and because it has taken into consideration some explicitly linguistic concepts in its design, some of which we have already discussed. In addition, it includes a very comprehensive section covering the major verb classes proposed by Levin (1993).

Grammatical Descriptions

We conclude by returning to the concept of linguistic inventory, and propose that a grammatical description, consisting of a set of assertions about the structure of a language, qualifies as an instance.

Suppose that all we know about Warumungu are the properties of the example in Figure 2. A description of that knowledge would consist of a set of assertions such as:

- *purrumu* expresses 'touch'.
- *rra* is a suffix for imperative mood.
- The third person singular object pronoun is phonologically null.
- *purrumurra* expresses 'touch it'.

On the other hand, someone who knows Warumungu well might be prepared to make much more general assertions such as the following, from Simpson (1998):

- Warumungu has six vowels: the three short vowels /i/, /a/, /u/, and their long counterparts.
- Primary stress is assigned to the first syllable of a word with two or more syllables.
- Warumungu has the following six classes of morphemes: nominals, preverbs, verb roots, bound pronouns, particles and suffixes.
- Warumungu words are formed by suffixation, reduplication, compounding, and template (for pronouns).
- Warumungu verb roots inflect for four forms of "associated motion": motion or direction toward the deictic center, motion or direction away from the deictic center, "setting out", and neutral.

The linguistic ontology, if properly constructed, should provide a metalanguage for expressing these assertions, enabling them to be compared with other grammatical descriptions of the same or of different languages. It would also support the ability to reason about a language, in order, for example, to determine whether a particular analysis of a particular example is consistent with a given grammatical description. Like the other linguistic inventories suggested here, grammatical descriptions can be very useful resources even if incomplete. They cannot however be inconsistent, so care must be exercised when merging two grammatical descriptions for the same language.

References

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

Gangemi A, Guarino N, Masolo C & Oltramari A (2001) Understanding top-level ontological distinctions. *Proceedings of IJCAI 2001 Workshop on Ontologies and Information Sharing*.

Guarino N & Welty C (2000) Ontological analysis of taxonomic relations. *Proceedings of ER-2000: The International Conference on Conceptual Modeling*.

Kendris C (1990) *501 Spanish Verbs*, 3rd ed. New York: Barron's.

Langendoen D T & Simons G F (1995) A rationale for the TEI recommendations for feature-structure markup. *Computers and the Humanities* 29, 191-209.

Lenat D & Guha R V (1990) *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*. Reading, MA: Addison-Wesley.

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

Maxwell M, Simons G F & Hayashi L (2002) Resources for morphology learning and evaluation: A morphological glossing assistant. *Proceedings of International Workshop on Resources and Tools for Field Linguistics*, Las Palmas, Spain.

Niles I & Pease A (2001) Toward a standard upper ontology. *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* 17, Special Issue on Ontologies and Agents.

Pollard C & Sag I A (1994) *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Shieber S M (1986) *An Introduction to Unification-Based Approaches to Grammar*. Stanford, CA: CSLI.

Simpson J (1998) Warumungu (Australian — Pama-Nyungan). *The Handbook of Morphology*, Spencer A & Zwicky A M, eds., 707-736. Oxford: Blackwell.