A Registry of Standard Data Categories for Linguistic Annotation

Nancy Ide[1] and Laurent Romary[2]
Department of Computer Science, Vassar College, Poughkeepsie, NY 12604-0520 USA
Equipe Langue et Dialogue, LORIA/INRIA, Vandoeuvre-lès-Nancy FRANCE
ide@cs.vassar.edu, romary@loria.fr

Abstract

In this paper we describe the most recent work within ISO TC37/SC 4, and in particular the development of a Data Category Registry (DCR) component of the Linguistic Annotation Framework. The DCR will contain a formally defined set of linguistic categories in common use within the language engineering community for reference and use in linguistically annotated resources. We outline the first proposals for creation and management of the DCR, as a solicitation for input from the community.

Introduction

Data associated with language resources are identified, collected, managed, and stored in a wide variety of formats and environments. Differences in approach among different language resources and individual system objectives inevitably lead to variations in data category definitions and data category names. The use of uniform data category names and definitions within the same resource domain (e.g., among terminological, lexicographical, text corpus, etc. resources), at least at the interchange level, contributes to system coherence and enhances the re-usability of data. Procedures for defining data categories in a given resource domain should also be uniform in order to ensure interoperability.

In this paper we describe the most recent work within ISO TC37/SC 4, and in particular the development of a Data Category Registry (DCR) component of the Linguistic Annotation Framework (LAF) (Ide and Romary, 2003; Ide, Romary, and Clergerie, 2003). The overall LAF architecture is based on the principle of separation of user annotation formats and data categories from the exchange/processing ("pivot") format on the one hand; and separation of the structure of annotations and their content on the other. The pivot format implements an abstract data model for annotations as a feature structure graph instantiated in XML, which reflects the internal structure of an annotation; annotation content is provided by attribute/value pairs attached to nodes in the graph. To support uniform definition of annotation content, the LAF architecture includes a Data Category Registry (DCR) containing a formally defined set of linguistic categories in common use within the language engineering community. The formally defined set of categories will have several functions: (1) it will provide a precise semantics for annotation categories that can be either used "off the shelf" by annotators or modified to serve specific needs; (2) it will provide a set of reference categories onto which scheme-specific names can be mapped; and (3) it will provide a point of departure for definition of variant, more precise, or entirely new data categories for use in language resource annotation.

We outline here the first proposals for creation and management of a DCR to support the creation and use of language resources for language engineering applications, as a solicitation for input from the community.

Background and Requirements

We define a *data category* as an elementary descriptor used in a linguistic annotation scheme. In feature structure terminology, data categories include both attributes (hereafter called *type descriptors*) such as SYNTACTIC CATEGORY and GRAMMATICAL GENDER, as well as a set of associated atomic *values* taken by such attributes, such as NOUN and FEMININE. In both cases we distinguish between the abstraction (concept) behind an attribute or value, and its realization as some string of characters or other object. Figure 1 provides an overview of these relationships. Whereas there is only one concept for a given attribute or value, there may be multiple instantiations.

type descriptor	*value*	
GENDER	MASCULINE FEMININE NEUTER	*conceptual dimension*
`gen`	`{m,f,n}`	*instantiation*
`genre`	`{masc, fem, neut}`	*instantiation*

Figure 1. Data category overview

The DCR under development within ISO TC37 SC4 is built around this fundamental concept/instance distinction. In principle, the DCR provides a set of reference concepts, while the annotator provides a *Data Category Specification* (DCS) that comprises a mapping between his or her scheme-specific instantiations and the concepts in the DCR. As such, the DCS provides documentation for the linguistic annotation scheme in question. The DCS for a given annotation document/s is included or referenced in any data exchange to provide the receiver with the

information required to interpret the annotation content or to map it to another instantiation. Semantic integrity is guaranteed by mutual reference to DCR concepts.

To serve the needs of the widest possible user community, the DCR must be developed with an eye toward multi-lingualism. The Data Category Registry will support multiple languages by providing the following:

- reference definitions for data categories in various languages;
- data element names for the data categories in various languages;
- description of usage in language-specific contexts, including definitions, usage notes, examples, and/or lists of values (e.g., GENDER takes the values *masculine*, *feminine* in French; *masculine, feminine, neuter* in German)

In addition, to both accommodate archival data and ensure the semantic integrity, a mapping of data categories instantiated in the DCR to categories and values in well-known projects and initiatives will be provided.

Managing the DCR

The DCR for language resources will be a reference for all the existing or future standards in TC37 related to data modeling or data interchange. At the moment, it is envisaged that ISO committee TC37 will implement a single, central data category registry covering all applications within the domain of language resource creation and use.

It is anticipated that management of the registry will not be fully centralized, but rather will implement a structure designed to bring together the right expertise within a subfield of linguistic resources and at the same time ensure coherence within the registry. Accordingly, the decision process that leads to the introduction or revision of a data category into the registry will be organized in two broad steps:

1. a *selection process* in which a committee of experts in a given domain, either identified by TC37 members or proposed by the relevant TC37 sub-committee chair, proposes a set data categories relevant for that domain. The selection process will implement a cycle of proposal, publication, solicitation of public comment, revision, and approval not unlike the usual ISO processes and the similar processes employed by the W3C.
2. a *harmonization process*, managed by a DCR board consisting of a group of experts and a chair appointed by the TC37 plenary. The role of the board role is to ensure the coherence of new proposals with the scope of the registry and the data categories it already contains.

The creation of a single global data category registry for all types of language resources treated within TC37 provides a unified view over the various applications of such a reference resource. However, for the purposes of both category creation and DCR access, the DCR will be organized according to *thematic views*, i.e. domains of activity, which include specialized subsets of the information in the registry. Given the on-going activities within TC37, we can envisage definable subsets of the DCR for at least the following: terminological data collection, various types of linguistic annotation (morpho-syntactic, syntactic, discourse level, etc.), lexical representation for both NLP-oriented and traditional lexicography, language resource metadata, and language codes.

Figure 2 illustrates the relationship between data category specifications and the DCR. The patterned cells correspond to individual DCS's. Some data categories are relevant to a single domain, while others are common to multiple domains: for example, *sense number* is probably specific to lexicographical resources, but linguistic categories such as *part of speech, grammatical gender, grammatical number,* etc. have wider application. Each thematic domain contributes all its data categories the global DCR, while at the same time identifying those data categories that it shares with other domains.

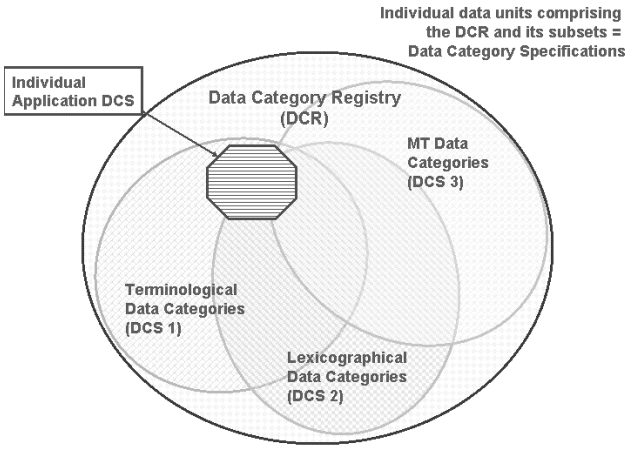

Figure 2. Relation of Data Category Selections to the DCR

The oval shapes in the Venn diagram represent DCS subsets. A smaller subset can be selected from the domain DCS for use in a given application, as represented by the octagon in Figure 2. Note that while some of the data categories contained in this subset are common to several different domains, this application is wholly contained within the DCS for terminological entries, so we can conclude that it is designed for use with a terminological application.

GettingStarted

Creation of the DCR can begin by considering or incorporating existing or developing ISO standards, including ISO 12620, which describes a set of reference data categories for terminology representation, and ISO 639, which is currently being extended to cover a much larger group than its previous version. Some of the data categories already defined in ISO 12620, for example, include general-purpose management data categories (e.g., SOURCE, RESPONSIBILITY, DATE, etc.) as well as linguistic categories (e.g., PART OF SPEECH), which can provide a base for extension. In addition, it should certainly be possible to utilize results from previous or existing projects such as EAGLES/ISLE to provide a base set of categories for consideration.

We intend to proceed cautiously, implementing categories that are widely used and relatively low-level, to ensure acceptance by the community. By building up slowly, the

DCR should eventually contain a wide range of data categories, with their complete history, data category description, and attendant metadata. It would then be possible to specify a DCS (see previous section) for different thematic domains and an ontology of relations among them. In the short term, it is likely not reasonable to define such an ontology until there is greater awareness and consensus at the international level. However, no choice should be made in the definition of the DCR that would hamper further work in this direction.

So far, we have defined a preliminary template for data category definitions to be used as an underlying model for the DCR (ISO TC37/SC 3 N488), which can also serve as a model for manipulation and transmission of proprietary data categories within the language engineering community. Figure 3 provides an overview of the general outline; the heart of a data category description is the *Conceptual Entry* section, which we define to include the following fields:

ENTRY IDENTIFIER used for interchange of data category

DEFINITION reference definition for the category, language and theory neutral to the extent possible.

EXPLANATION additional information about the data category not relevant in a definition (e.g. more precise linguistic background for the use of the data category);

EXAMPLE illustration of use of the category, excluding language specific usages (documented elsewhere)

SOURCE may refine definition, explanation, or example to indicate the source from which the corresponding text has been borrowed or adapted.

STATUS may refine definition to indicate approval, acceptability, or applicability in a given context

PROFILE relates the current data category to one or several views (e.g. Morpho-syntax, Syntax, Metadata, Language description, etc.)

CONCEPTUAL RANGE relates the category to the set of possible values (expressed as a list of data categories). A datatype may be provided instead of a list of values

NOTE additional information excluding technical information that would normally be described within explanation

BROADER CONCEPT GENERIC pointer to a more general data category (e.g., from Common noun to Noun)

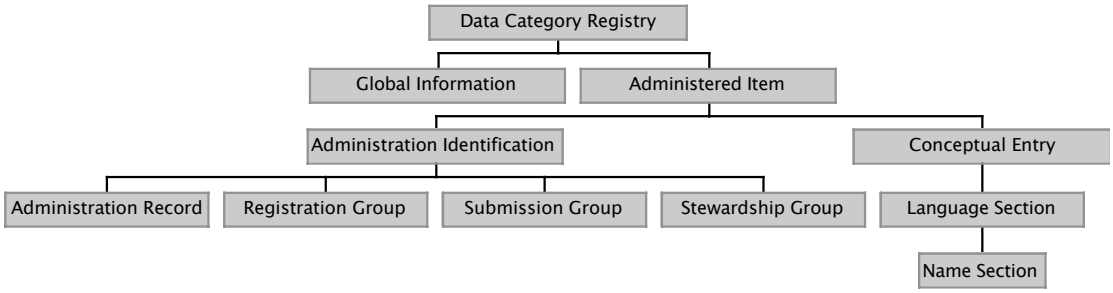

Figure 3. Overview of DCR structure.

Using the DCR

The purpose of the DCR is to promote greater usability and reusability of annotated language resources and increased semantic integrity for information in annotation documents by providing a set of formally-defined reference categories. "Formal definition" in this context includes natural language definitions for each category accompanied by specification of the possible values each category may take. At present, we envision instantiation of the DCR as a simple database in which each entry is either a type descriptor or value. Data categories will be referenced either by the DCR entry identifier, or, since the DCR will be publicly available on-line, via a URI.

Note that this simple instantiation of the DCR makes no distinction in terms of representation between type descriptors and values; each is considered as a data category and provided with an entry identifier for reference. Only minimal constraints on their use in an annotation are specified--i.e., constraints on descriptor/value combinations given in the descriptor entry. The broader structural integrity of an annotation is provided by placing constraints on nodes in the annotation graph (as defined in the LAF architecture) with which a given category can be associated. For example, the structural graph for a syntactic constituency analysis would consist of a hierarchy of typed nodes corresponding to the non-terminals in the grammar, with constraints on their embedding, and with which only appropriate descriptor/value pairs may be associated. Node types (e.g., NP, VP) as well as associated grammatical information (e.g., tense, number) may all be specified with data categories drawn from the DCR.

A more formal specification of data categories could be provided using mechanisms such as RDF Schema (RDFS) and the Ontology Web Language (OWL) to formalize the properties and relations associated with data categories. For example, consider the following RDF Schema fragment:

```
<rdfs:Class rdf:about="#Noun">
  <rdfs:label>Noun</rdfs:label>
  <rdfs:comment>Class for
          nouns</rdfs:comment>
</rdfs:Class>
<rdfs:Property rdf:about="#number">
   <rdfs:domain
      rdfs:resource="Noun"/>
   <rdfs:range
      rdf:resource="rdfs:#Literal"/>
</rdfs:Property>
```

This fragment defines a class of objects called " "Noun" that has a property "number". Note that the schema defines the classes but does not instantiate objects belonging to the class; instantiation may be accomplished directly in the annotation file, as follows (for brevity, the following examples assume appropriate namespace declarations specifying the URIs of schema and instance declarations):

```
<Noun rdf:about="Mydoc#W1">
     <number rdf:value="Plural"/>
</Noun>
```

where "Mydoc#W1" is the URI of the word being annotated as a noun. Alternatively, the DCR could contain instantiations of basic data elements, specifying values for properties, which can be referenced directly in the annotation:

```
<Noun rdf:ID="NMP">
    <number rdf:value="plural"/>
</Noun>
```

The annotation file could then reference the pre-defined instance as follows:

```
<rdf:Description rdf:about="myDoc#W1">
    <POS rdf:resource="categories#NMS"/>
</rdf:Description>
```
[1]

An RDFS/OWL specification of data categories would enable greater control over descriptor/value use and also allow for the possibility of inferencing over annotations. However, it would also demand definition of a precise hierarchy of linguistic categories and a distinction between classes (objects) and properties that could place unwanted constraints on annotation form and content. Therefore, any such specification of data categories is left to the annotator, at least for the time being.

It is anticipated that many annotators will use their own category names and values and provide a mapping to DCR categories. The DCR will include an XML template for specifying this mapping, as well as for defining variants and new descriptor/value pairs.

[1] In these examples, NUMBER is given literal values. However, with OWL it is possible to restrict the range of possible values by enumeration.

Conclusion

The goal of the DCR is not to impose a specific set of categories, but rather to ensure that the semantics of data categories included in annotations (whether they exist in the DCR or not) are well-defined and understood, by gathering together (and where necessary, harmonizing) existing categories in use by the language technology community as a resource for the annotation of linguistic data. It is possible that several different instantiations of the same category (e.g., noun) and/or different schemas describing the same phenomenon will exist in the DCR, to be used as desired by annotators. The aim is not to prescribe, but rather to move toward, commonality in annotation content, which is becoming more and more essential as annotated language data is increasingly distributed over multiple sites and accessible via the web.

The DCR can succeed only with the input of the language engineering and computational linguistics communities. We invite feedback and comments on design drafts for the DCR and the Linguistic Annotation Framework in general. For general information on the work of the ISO committee on language resources, consult the ISO TC37/SC4 website (http://www.tc37sc4.org).

References

Ide, N. & Romary, L. (2003). Outline of the International Standard Linguistic Annotation Framework. Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right, Sapporo, 1-5.

Ide, N., Romary, L., & de la Clergerie, E. (2003). International Standard for a Linguistic Annotation Framework. Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology, Edmunton.