International Standard for a Linguistic Annotation Framework

Nancy Ide
Dept. of Computer Science
Vassar College
Poughkeepsie,
New York 12604-0520
USA
ide@cs.vassar.edu

Laurent Romary
Equipe Langue et Dialogue
LORIA/INRIA
Vandoeuvre-lès-Nancy
FRANCE
romary@loria.fr

Eric de la Clergerie
INRIA Rocquencourt, BP 105
78153 Le Chesnay cedex
FRANCE
Eric.De_La_Clergerie@inria.fr

Abstract

This paper describes the outline of a linguistic annotation framework under development by ISO TC37 SC WG1-1. This international standard will provide an architecture for the creation, annotation, and manipulation of linguistic resources and processing software. The outline described here results from a meeting of approximately 20 experts in the field, who determined the principles and fundamental structure of the framework. The goal is to provide maximum flexibility for encoders and annotators, while at the same time enabling interchange and re-use of annotated linguistic resources.

1 Introduction

Language resources are bodies of electronic language data used to support research and applications in the area of natural language processing. Typically, such data are enhanced (annotated) with linguistic information such as morpho-syntactic categories, syntactic or discourse structure, co-reference information, etc.; or two or more bodies may be aligned for correspondences (e.g., parallel translations, speech signal and transcription).

Over the past 15-20 years, increasingly large bodies of language resources have been created and annotated by the language engineering community. Certain fundamental representation principles have been widely adopted, such as the use of stand-off annotation, use of XML, etc., and several attempts to provide generalized annotation mechanisms and formats have been developed (e.g., XCES, annotation graphs). However, it remains the case that annotation formats often vary considerably from resource to resource, often to satisfy constraints imposed by particular processing software. The language processing community has recognized that commonality and interoperability are increasingly imperative to enable sharing, merging, and comparison of language resources.

To provide an infra-structure and framework for language resource development and use, the International Organization for Standardization (ISO) has formed a sub-committee (SC4) under Technical Committee 37 (TC37, Terminology and Other Language Resources) devoted to Language Resource Management. The objective of ISO/TC 37/SC 4 is to prepare international standards and guidelines for effective language resource management in applications in the multilingual information society. To this end, the committee is developing principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes. The focus of the work is on data modeling, markup, data exchange and the evaluation of language resources other than terminologies (which have already been treated in ISO/TC 37). The worldwide use of ISO/TC 37/SC 4 standards should improve information management within industrial, technical and scientific environments, and increase efficiency in computer-supported language communication.

At present, language professionals and standardization experts are not sufficiently aware of the standardization efforts being undertaken by ISO/TC 37/SC 4. Promoting awareness of future activities and rising problems, therefore, is crucial for the success of the committee, and will be required to ensure widespread adoption of the standards it develops. An even more critical factor for the success of the committee's work is to involve, from the outset, as many and as broad a range of potential users of the standards as possible.

Within ISO/TC 37/SC 4, a working group (WG1-1) has been established to develop a Linguistic Annotation Framework (LAF) that can serve as a basis for harmonizing existing language resources as well as developing new ones. In order to ensure that the framework is developed based on the input and consensus of the research community, a group of experts[1] was convened on November 21-22, 2002, at Pont-à-Mousson, France, to lay out the overall structure of the framework. .In this paper, we outline the conclusions from this meeting, and solicit the input of other members of the community to inform its further development.

2 Background and rationale

The standardization of principles and methods for the collection, processing and presentation of language resources requires a distinct type of activity. Basic standards must be produced with wide-ranging applications in view. In the area of language resources, these standards should provide various technical committees of ISO, IEC and other standardizing bodies with the groundwork for building more precise standards for language resource management.

The need for harmonization of representation formats for different kinds of linguistic information is critical, as resources and information are more and more frequently merged, compared, or otherwise utilized in common systems. This is perhaps most obvious for processing multi-modal information, which must support the fusion of multimodal inputs and represent the combined and integrated contributions of different types of input (e.g., a spoken utterance combined with gesture and facial expression), and enable multimodal output (see, for example, Bunt and Romary, 2002). However, language processing applications of any kind require the integration of varieties of linguistic information, which, in today's environment, come from potentially diverse sources. We can therefore expect use and integration of, for example, syntactic, morphological, discourse, etc. information for multiple languages, as well as information structures like domain models and ontologies.

We are aware that standardization is a difficult business, and that many members of the targeted communities are skeptical about imposing any sort of standards at all. There are two major arguments against the idea of standardization for language resources. First, the diversity of theoretical approaches to, in particular, the annotation of various linguistic phenomena suggests that standardization is at least impractical, if not impossible. Second, it is feared that vast amounts of existing data and processing software, which may have taken years of effort and considerable funding to develop, will be rendered obsolete by the acceptance of new standards by the community. Recognizing the validity of both of these concerns, WG1-1 does not seek to establish a single, definitive annotation scheme or format. Rather, the goal is to provide a framework for linguistic annotation of language resources that can serve as a reference or pivot for different annotation schemes, and which will enable their merging and/or comparison. To this end, the work of WG1-1 includes the following:

- o analysis of the full range of annotation types and existing schemes, to identify the fundamental structural principles and content categories;
- o instantiation of an abstract format capable of capturing the structure and content of linguistic annotations, based on the analysis in (1);
- o establishment of a mechanism for formal definition of a set of reference content categories which can be used "off the shelf" or serve as a point of departure for precise definition of new or modified categories.
- o provision of both a set of guidelines and principles for developing new annotation schemes and concrete mechanisms for their implementation, for those who wish to use them.

By situating all of the standards development squarely in the framework of XML and related standards such as RDF, DAML+OIL, etc., we hope to ensure not only that the standards developed by the committee provide for compatibility with established and widely accepted web-based technologies, but also that transduction from legacy formats into XML formats conformant to the new standards is feasible.

3 General requirements for a linguistic annotation framework

The following general requirements for a linguistic annotation framework were identified by the group of experts at Pont-à-Mousson:

Expressive adequacy

The framework must provide means to represent all varieties of linguistic information (and possibly also other types of information). This includes representing the full range of information from the very general to information at the finest level of granularity.

[1] Participants: Nuria Bel (Universitat de Barcelona), David Durand (Brown University), Henry Thompson (University of Edinburgh), Koiti Hasida (AIST Tokyo), Eric De La Clergerie (INRIA), Lionel Clement (INRIA), Laurent Romary (LORIA), Nancy Ide (Vassar College), Kiyong Lee (Korea University), Keith Suderman (Vassar College), Aswani Kumar (LORIA), Chris Laprun (NIST), Thierry Declerck (DFKI), Jean Carletta (University of Edinburgh), Michael Strube (European Media Laboratory), Hamish Cunningham (University of Sheffield), Tomaz Erjavec (Institute Jozef Stefan), Hennie Brugman (Max-Planck-Institut für Psycholinguistik), Fabio Vitali (Universite di Bologna), Key-Sun Choi (Korterm), Jean-Michel Borde (Digital Visual), Eric Kow (LORIA).

Media independence

The framework must handle all potential media types, including text, audio, video, image, etc. and should, in principle, provide common mechanisms for handling all of them. The framework will rely on existing or developing standards for representing multi-media.

Semantic adequacy

- Representation structures must have a formal semantics, including definitions of logical operations
- There must exist a centralized way of sharing descriptors and information categories

Incrementality

- The framework must provide support for various stages of input interpretation and output generation.
- The framework must provide for the representation of partial/under-specified results and ambiguities, alternatives, etc. and their merging and comparison.

Uniformity

Representations must utilize same "building blocks" and the same methods for combining them.

Openness

The framework must not dictate representations dependent on a single linguistic theory.

Extensibility

The framework must provide ways to declare and interchange extensions to the centralized data category registry.

Human readability

Representations must be human readable, at least for creation and editing.

Processability (explicitness)

Information in an annotation scheme must be explicit—that is, the burden of interpretation should not be left to the processing software.

Consistency

Different mechanisms should not be used to indicate the same type of information.

To fulfill these requirements, it is necessary to identify a consistent underlying *data model* for data and its annotations. A data model is a formalized description of the data objects (in terms of composition, attributes, class membership, applicable procedures, etc.) and relations among them, independent of their instantiation in any particular form. A data model capable of capturing the structure and relations in diverse types of data and annotations is a pre-requisite for developing a common corpus-handling environment: it impacts the design of annotation schema, encoding formats and data architectures, and tool architectures.

As a starting assumption, we can conceive of an annotation as a one- or two-way link between an annotation object and a point (or a list/set of points) or span (or a list/set of spans) within a base data set. Links may or may not have a semantics--i.e., a type--associated with them. Points and spans in the base data may themselves be objects, or sets or lists of objects. We make several observations concerning this assumption:

- the model assumes a fundamental linearity of objects in the base,[2] e.g., as a time line (speech); a sequence of characters, words, sentences, etc.; or pixel data representing images;
- the *granularity* of the data representation and encoding is critical: it must be possible to uniquely point to the smallest possible component (e.g., character, phonetic component, pitch signal, morpheme, word, etc.);
- an annotation scheme must be mappable to the structures defined for annotation objects in the model;
- an encoding scheme must be able to capture the object structure and relations expressed in the model, including class membership and inheritance, therefore requiring a sophisticated means to specify linkage within and between documents;
- it is necessary to consider the logistics of identifying spans by enclosing them in start and end tags (thus enabling hierarchical grouping of objects in the data itself), vs. explicit addressing of start and end points, which is required for read-only data;
- it must be possible to represent objects and relations in some (fairly straightforward) form that prevents information loss;
- ideally, it should be possible to represent the objects and relations in a variety of formats suitable to different tools and applications.

ISO TC37/SC 4's goal is to develop a framework for the design and implementation of linguistic resource formats and processes in order to facilitate the exchange of information between language processing modules. A well-defined representational framework for linguistic information will also provide for the specification and comparison of existing application-specific representations and the definition of new ones, while ensuring a level of interoperability between them. The framework should allow for variation in annotation schemes while

[2] Note that this observation applies to the *fundamental* structure of stored data. Because the targets of a relation may be either individual objects, or sets or lists of objects, information with more than one dimension is accommodated.

at the same time enabling comparison and evaluation, merging of different annotations, and development of common tools for creating and using annotated data. For this purpose we envisage a common "pivot" format based on a data model capable of capturing all types of information in linguistic annotations, into and out of which site-specific representation formats can be transduced. This strategy is similar to that adopted in the design of languages intended to be reusable across platforms, such as Java. The pivot format must support the communication among all modules in the system, and be adequate for representing not only the end result of interpretation, but also intermediate results.

4 Terms and definitions

The following terms and definitions are used in the discussion that follows:

Annotation: The process of adding linguistic information to language data ("annotation of a corpus") or the linguistic information itself ("an annotation"), independent of its representation. For example, one may annotate a document for syntax using a LISP-like representation, an XML representation, etc.

Representation*:* The format in which the annotation is rendered, e.g. XML, LISP, etc. independent of its content. For example, a phrase structure syntactic annotation and a dependency-based annotation may both be represented using XML, even though the annotation information itself is very different.

Types of Annotation*:* We distinguish two fundamental types of annotation activity:

1. *segmentation* : delimits linguistic elements that appear in the primary data. Including
   - o continuous segments (appear contiguously in the primary data)
   - o super- and sub-segments, where groups of segments will comprise the parts of a larger segment (e.g., a contiguous word segments typically comprise a sentence segment)
   - o discontinuous segments (linking continuous segments)
   - o landmarks (e.g time stamps) that note a point in the primary data

   In current practice, segmental information may or may not appear in the document containing the primary data itself. Documents considered to be *read-only,* for example, might be segmented by specifying byte offsets into the primary document where a given segment begins and ends.

2. *linguistic annotation:* provides linguistic information about the segments in the primary data, e.g., a morpho-syntactic annotation in which a part of speech and lemma are associated with each segment in the data. Note that the identification of a segment as a word, sentence, noun phrase, etc. also constitutes linguistic annotation. In current practice, when it is possible to do so, segmentation and identification of the linguistic role or properties of that segment are often combined (e.g., syntactic bracketing, or delimiting each word in the document with an XML tag that identifies the segment as a word, sentence, etc.).

Stand-off annotation: Annotations layered over a given primary document and instantiated in a document separate from that containing the primary data. Stand-off annotations refer to specific locations in the primary data, by addressing byte offsets, elements, etc. to which the annotation applies. Multiple stand-off annotation documents for a given type of annotation can refer to the same primary document (e.g., two different part of speech annotations for a given text). There is no requirement that a single XML-compliant document may be created by merging stand-off annotation documents with the primary data; that is, two annotation documents may specify trees over the primary data that contain overlapping hierarchies.

5 Design principles

The following general principles will guide the LAF development:

- o The data model and document form are distinct but mappable to one another
- o The data model is parsimonious, general, and formally precise
- o The data model is built around a clear separation of structure and content
- o There is an inventory of logical operations supported by the data model, which define its abstract semantics
- o The document form is largely under user control
- o The mapping between the flexible document form and data model is via a rigid dump-format
- o The mapping from document form to the dump format is documented in an XML Schema (or the functional equivalent thereof) associated with the document
- o Mapping is operationalized *either* via schema-based data-binding process *or* via schema-derived stylesheet mapping between the user document and the dump-format document.

- It must be possible to isolate specific layers of annotation from other annotation layers or the primary (base) data; i.e., it must be possible to create a format using stand-off annotation
- The dump format must be designed to enable stream marshalling and unmarshalling

The overall architecture of LAF as dictated by these principles is given in Figure 1. The left side of the diagram represents the user-defined document form, and is labeled "human" to indicate that creation and editing, of the resource is accomplished via human interaction with this format. This format should, to the extent possible, be human readable. We will support XML for these formats (e.g., by providing style sheets, examples, etc.) but not disallow other formats. The right side represents the dump format, which is machine processable, and may not be human readable as it is intended for use only in processing. This format will be instantiated in XML.

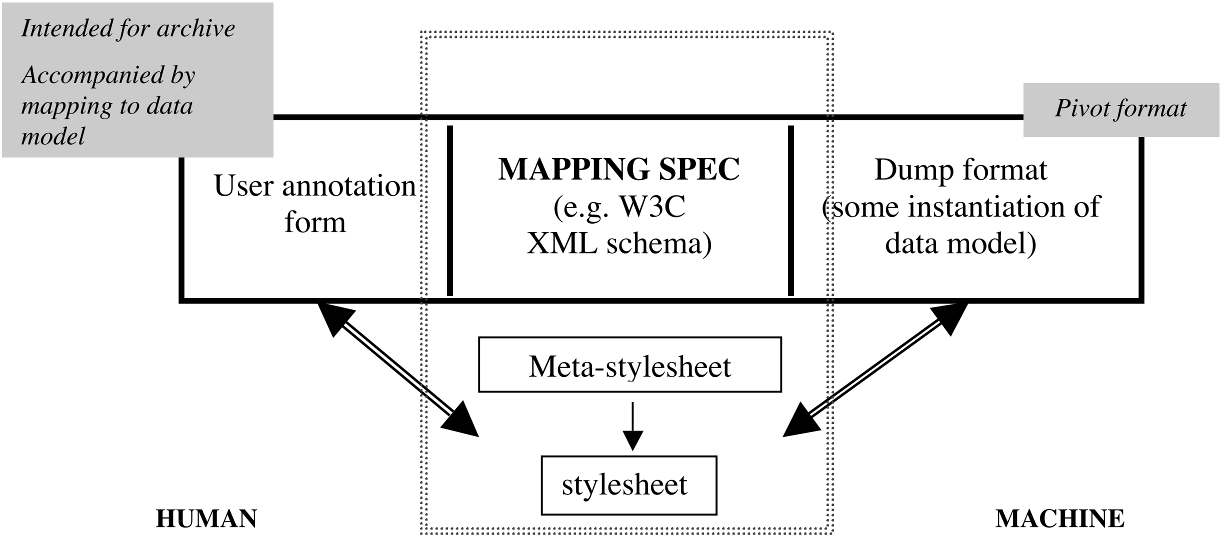
Intended for archive
Accompanied by mapping to data model
Pivot format
User annotation form
MAPPING SPEC (e.g. W3C XML schema)
Dump format (some instantiation of data model)
Meta-stylesheet
stylesheet
HUMAN
MACHINE

Figure 1. Overall LAF architecture

6 Practice

The following set of practices will guide the implementation of the LAF:

- The data model is essentially a feature structure graph with a moderate admixture of algebra (e.g. disjunction, sets), grounded in *n*-dimensional regions of primary data and literals.
- The dump format is isomorphic to the data model.
- Semantic coherence is provided by a registry of features in an XML-compatible format (e.g., RDF), which can be used directly in the user-defined formats and is always used with the dump format.
- Resources will be available to support the design and specification of document forms, for example:
    - XML Schemas in several normal forms based on type definitions and abstract elements that can be exploited via type derivation and/or substitution group;
    - XPointer design-patterns with standoff semantics;
    - Schema annotations specifying mapping between document form and data model;
    - Meta-stylesheet for mapping from annotated XML Schema to mapping stylesheets;
    - Data-binding stylesheets with language-specific bindings (e.g. Java).
- Users may define their own data categories or establish variants of categories in the registry. In such cases, the newly defined data categories will be formalized using the same format as definitions available in the registry, and will be associated with the dump format.
- The responsibility of converting to the dump format is on the producer of the resource.
- The producer is responsible for documenting the mapping from the user format to the data model.

- The ISO working group will provide test suites and examples following these guidelines:
    - The example format should illustrate use of data model/mapping
    - The examples will show both the left (human-readable) and right (machine processable) side formats
    - Examples will be provided that use existing schemes

7 Discussion

The framework outlined in the previous section provides for the use of any annotation format consistent with the feature structure-based data model that will be used to define the pivot format. This suggests a future scenario in which annotators may create and edit annotations in a proprietary format, transduce the annotations using available tools to the pivot format for interchange and/or processing, and if desired, transduce the pivot form of the annotations (and/or additional annotation introduced by processing) back into the proprietary format. We anticipate the future development of annotation tools that provide a user-oriented interface for specifying annotation information, and which then generate annotations in the pivot format directly. Thus the pivot format is intended to function in the same way as, for example, Java byte code functions for programmers, as a universal "machine language" that is interpreted by processing software into an internal representation suited to its particular requirements. As with Java byte code, users need never see or manipulate the pivot format; it is solely for machine consumption.

Information units or *data categories* provide the semantics of an annotation. Data categories are the most theory and application-specific part of an annotation scheme. Therefore, LAF includes a Data Category Registry to provide a means to formally define data categories for reference and use in annotation. To make them maximally interoperable and consistent with existing standards, RDF schemas can be used to formalize the properties and relations associated with each data category. The RDF schema ensures that each instantiation of the described objects is recognized as a sub-class of more general classes and inherits the appropriate properties. Annotations will reference the data categories via a URL identifying their instantiations in the Data Category Registry itself. The class and sub-class mechanisms provided in RDFS and its extensions in OWL will also enable creation of an ontology of annotation classes and types.

A 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 or more precise categories. Thus the overall goal of the Data Category Registry 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 Registry or not) are well-defined and understood.

The data model that will define the pivot format must be capable of representing all of the information contained in diverse annotation types. The model we assume is a feature structure graph for annotation information, capable of referencing *n*-dimensional regions of primary data as well as other annotations. The choice of this model is indicated by its almost universal use in defining general-purpose annotation formats, including the Generic Modeling Tool (GMT) (Ide & Romary, 2001, 2002) and Annotation Graphs (Bird & Liberman, 2001). The XML-based GMT could serve as a starting point for defining the pivot format; its applicability to diverse annotation types, including terminology, dictionaries and other lexical data (Ide, *et al*., 2000), morphological annotation (Ide & Romary, 2001a; 2003) and syntactic annotation (Ide & Romary, 2001b) demonstrates its generality. As specified by the LAF architecture, the GMT implements a feature structure graph, and exploits the hierarchical structure of XML elements and XML's powerful inter- and intra-document pointing and linkage mechanisms for referencing both "raw" and XML-tagged primary data and its annotations.

The provision of development resources, including schemas, design patterns, and stylesheets, will enable annotators and software developers to immediately adapt to LAF. Example mappings, e.g., for XCES-encoded annotations, will also be provided.

8 Conclusion

In this paper we describe the Linguistic Annotation Framework under development by ISO TC37/SC 4 WG1-1, as defined by a group of experts convened at a workshop in Pont-à-Mousson, France, in late 2002. Its design is intended to allow for, on the one hand, maximum flexibility for annotators, and. on the other, processing efficiency and reusability. This is accomplished by separating user annotation formats from the exchange/processing format. This separation also ensures that pre-existing annotations are compatible with LAF.

ISO TC37/SC4 is just beginning its work, and will use the general framework discussed in the preceding sections as its starting point. However, the work of the committee will not be successful unless it is accepted by the language processing community. To ensure widespread acceptance, it is critical to involve as many representatives of the community in the development of the standards as possible, in order to ensure that all needs are addressed. This paper serves as a call for participation to the language processing community; those interested should contact the TC 37/SC 4 chairman (Laurent Romary: romary@loria.fr). For general information, consult the ISO TC37/SC4 website (http://www.tc37sc4.org).

References

Bird, S. & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication,* 33:1-2, 23-60.

Bunt. H. & Romary, L. (2002). Towards Multimodal Content Representation. *Proceedings of the Workshop on International Standards for Terminology and Language Resource Management*, Las Palmas.

Ide, N. & Romary, L. (2001a). Standards for Language Resources, *IRCS Workshop on Linguistic Databases*, Philadelphia, 141-49.

Ide, N. & Romary, L. (2001b). A Common Framework for Syntactic Annotation. *Proceedings of ACL'2001,* Toulouse, 298-305.

Ide, N., Kilgarriff, A., & Romary, L. (2000). A Formal Model of Dictionary Structure and Content. *Proceedings of Euralex 2000,* Stuttgart, 113-126.

Ide, N. & Romary, L. (2003). Encoding Syntactic Annotation. In Abeillé, A. (ed.). *Treebanks: Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer Academic Publishers (in press).