

Language Resource Management
Descriptors and Mechanisms for Language Resources

File ID wg1N19.doc (49 ko)
wg1N19.pdf (23 ko)

Title : ISO Meeting workgroup notes

Editor(s) : Keyong Lee

Source : WG 1

Project number : N/A

Status : Contribution -

Date : 2002-11-22

Agenda / Action : For consideration, discussion and comment

References :

Mr. Key-Sun Choi - SC4 Secretary – KORTERM - 373-1 Kusong-Dong Yusong-gu - Taejon 305-701 - Korea
82 42 869 35 25 – fax: 82 42 867 35 65 - kschoi@cs.kaist.ac.kr – <http://korterm.kaist.ac.kr>

Any question about SC4/WG1 documents registration, maintenance and distribution must be forwarded to Afnor
France preferably to: sylvie.arbouy@afnor.fr

workgroup notes

- **Formats WG**
- **Query Languages**
- **Data models (commonalities)**
- **Reason for work**

To share data between groups to share data across tools, and to be able to use tools from other places on one's data. even things like annotation manuals might eventually be part of this group. 3 kinds of interoperability: levels: morphosyntax/ grammar basic structure: that we have way to describe the same structures consistently between projects and tools. (horrible example: the xerox FSA notation). descriptors: At the third layer, we would like to be sure that a lemma is the same thing across projects: What is the ontology of the objects that we are describing by annotation. Question: how does this relate to the work of the terminology working group? For machine learning you just want lots of data, but aren't too picky about many details. For research, one may want to compare what is happening at different projects, so that a framework like this should not constrain research groups to particular practices, but rather enable description. Industry wants consistency in the outputs.

- **scope of the work**

TEI: annotation needs for scholars -- anyone doing some commenting of a text. ISO: Language engineering: precise descriptors intended for computer processing rather than human consumption. Transparency and debugability. Are these issues that impose requirements on the formats? Verification of data quality. The issue is not who is creating annotations, but the intended consumer of those annotations. We consider representation of resources that are not annotations on primary texts, but are secondary resources. We need not work on this swiftly but wait for others to pave the way. We should define what the minimal detail is for annotation of a particular type.

- **initial questions**

What resources are we annotating, and what aspects of language are we annotating on those resources?

- **primary data (what is being annotated)**

what kinds of resources themselves: signals (audio, images, video) texts (transcripts, documents) structured texts: some texts already have "things that would not change in future annotation". Are we dealing with the relation of non-electronic formats to their "transcriptions"? Answering no, this is not in scope. The TEI deals with things like this, and we thus do. by transitivity of use of TEI resources.

- **resources with a purpose**

Data sets with a purpose: to explore conversational strategies, dialogue, etc. Language engineering preservation of endangered languages.

- **what is an annotation**

An annotation is a structured representation of an interpretation of a part of a text. Types 1 codings are mutually exclusive and exhaustive sets of categories that apply to spans of data. 2. human-readable comments attached to spans in the primary data. Our definition, based partially on the above: an annotation is a homogeneous set of statements segmenting and qualifying the primary data or a fragment of it (with reference to an agreed set of categories) note that data is primary with respect to an annotation if that annotation qualifies it.

- **what levels of annotation**

morphosyntax treebanks coreference annotation transcription -- primary data, not annotation purpose of MPEG-7 is to interpret propositional content of the document. They refer to many of these levels, as well as quantifier scopes, strict/sloppy, etc. These are all: prosody Semantic annotations how detailed do we go in the internal structures of this kind of annotation. discourse not mature (parts are parts are not) gesture (not mature) fluent/disordered speech background noise: don't process this tape!

- **morphosyntax**
- **treebanks**
- **coreference annotation**

- **prosody**
- **discourses**
- **gesture**
- **semantics**
- **Data structures used by annotations**
 - treesfeature structures
 - **trees**
 - two schools of tree interpretation:old, based strongly on context free grammar.new, based on translating lambda calculus to a tree representation.
 - **Feature structures**
 - two types in use:non-typed FStyped FS (where attributes have strictly constrained hierarchies). Linking mechanisms and sharing are very heavily used in typed FS.Uses: Unification and compatibility (subsumtion?)
 - **Graphs**
 - **Automata**
 - **data management**
 - **History should include management information**
 - All data should be able to be described as to where it came from, and under what authority.What tool and version was used to create the data.If human created, what person or organization created a particular piece of data.
 - **version control, and history of annotations**
 - **Lexical resources**
 - This is not primary data.the data needs frequently grow for consumers of lexical data; Are there too many kinds of data to record, so that standardization is impossible at this time?what about things like word-net which is halfway between lexivcal and ontological information.
- **Additional issues**
 - **Official status**
 - Fabio: legal databases have several kinds of status depending on who makes the annotation and under what authority
 - **multimodality**
 - Annotations depend on the modality of the data that is being annotated. For instance, if we have utterances, gaze, gestures, other non-linguistic data. This is different from multimedia.Indexical expressions for instance, may depend on gestures: "this one is better than that one". Contextual information.This is an area where we must be careful not to duplicate work already going on in multimedia working groups within the ISO. This group can provide input to ISO work in these areas.
 - **temporal information**
 - Time information may percolate from primary data to higher levels of annotation.
 - **processes**
 - does qapplciation of processies like unification and compatibility actually make a difference to the representation of structures like trees and Feature structures.No.
- **priorities**
 - pragmatic prios: something exists, so we should just use it. {the easier and more useful, the better}forcing priorities: there's something missing that blocks progress, so we need to create a standard to fill the gap.abstract data models: descriptions of the data types that are used in various annotation systems/tools.two kinds of model:• mathematics of annotations.mathematical models allow comparison.• meaning of levels (morphosyntax, etc.)danger of argument infinitely prolongedallows

more meaningful definitions of categories, structures (e.g. POS).extensive definitions in data sets can be useful, if grounded in a formal mathematical object.What to annotate?Why to annotate?How to annotate?Which of these three is key in priority setting?

