

ISO TC 37 SC4 N472

Minutes
May 25, 2008
Marrakesh
DRAFT

Minutes on the ISO TC 37 SC Meeting on May 25, 2008 are included below.

Nancy Ide	Linguistic Annotation Framework
Harry Bundt	Semantic Annotation Framework Part 2: Dialog Acts
Hasida Koiti	TDG6: Language Resource Ontologies
Kiyong Lee	Feature Structure Document
Kiyong Lee	Pre-DIS 24617-1 SemAF/Time.
Daan Broder	Proposed standard for References and Citations
Paul Buitelaar	Lexical Markup Framework and LingINFO
Nancy Ide and Alex Fang	Harmonization
Monica Monachini	Thematic Domain Group 7

Minutes for the first three presentations are provided by Alex Fang. Minutes for the remaining presentations are provided by Jennifer DeCamp. Minutes are reviewed and edited by Kiyong Lee.

Morning Sessions

Alex Fang

Linguistic Annotation Framework Nancy Ide

- I Principles
 - n Separation of data and annotation
 - n Dump format (mappable to other formats)
- I Slow revolution
 - n Model development
 - n Consideration of processing needs
 - n Proof of concept instantiation in the ANC
 - u Transduction of several different annotation thypes and formats to laf format
 - u Api to merge, transducer to other formats
- I Laf status
 - n Reduced fs specification
 - n Final xml forat /schema
 - u GrAF: Graph annotation format
 - n Mapping rules and examples
 - n Coordination with UIMA
 - n Header specification including information about annotation, similar to UIMA type definition
- I Basic model
 - n Annotation content represented by feature structures

- n Referential structure represented as a directed acyclic graph (DAG), which enables exploitation of well-understood graph traversal
- l Primary data
 - n No annotations for primary data
 - u Read only
 - u Modifications can be regarded as annotations
 - n Base segmentation of the primary data
 - u Compatible annotations, those that can be merged etc, use common base segmentation
- l Segmentation
 - n Set of disjoint edges over primary data, vertices, located between nodes
 - n Edge for hyphens used for conjoined elements, twenty-four (twenty plus four), how about anti-social?
 - n Edges are labeled and unlabelled edges default to pointing to an unordered list of constituents.
- l Advantages
 - n Easily applied
 - n Finding sub graphs
 - n Identify connected components
- l Annotation sets
 - n As
 - n Nodeset
 - n Edgeset
- l FS
 - n Each node as a feature value
- l Head
 - n Revision history
 - n Source
 - n URL of annotation documentation
 - n Reference to DCR/UIMA type system
 - n Type system specification
- l Harmonisation
 - n Interproject coordination committee
 - n Mapping of each format
 - u Proof of concept
 - u Reveal inconsistencies, problem
- l Thierry: Use of ontological resource for annotation
- l Keith: This should sit on top of LAF, separate from each other.
- l Nancy: Keep it simple and semantics elsewhere outside LAF.
- l Laurent: Other graph people (they use Nancy's system). More uniform names as in TEI?
 - Nancy: No, convention in graph community. Names have been chosen for good conformance already. Arc for instance is not often used in the community.
- l Laurent: send doc to Chris Cox to make it more ISO looking. Agreed by Nancy.
- l Nancy would need two more days to submit the draft, say June 15 2008.
- l Laurent and Kiyong have spent time talking with other groups for terminology

harmonization.

- | Keith has released API in public domain.
- | Laurent recommended stabilised api.
- | Nancy questioned data category registry's availability.
- | Laurent recommended bibliography for laf. Send it off for CD ballot immediately but have the doc proofread first.

Semantic annotation framework part 2: dialogue act

ISO/TC37/SC4 N442 rev00

Harry Bunt

- | Semaf related to Nancy's work on laf
- | Purpose and justification
 - n Widely used
 - n Particular useful for function and intension, design of dialogue systems
- | Acts
 - n Questions
 - n Statement
 - n Answer
 - n Confirmation
 - n Request
 - n Instruct
 - n Promise
 - n Acknowledgement
 - n Greeting
- | Past approaches
 - n Tranins, map taks, verbmobil, damsl, swbd-damsl, conconut...
 - u Underlying approaches to dialogue modeling
 - u Definitions of basic concepts
 - u Level of granularity
 - u And mutually inconsistent terminology
 - u Lack of solid foundations of definitions and multidimensionality
 - u Lack of interoperability
- | ISO approach
 - n Preparatory studies in tdg3, lirics
 - n Focus: to best support theannotaiton of dialogues with dialogue act information in an empricially and theoretically well founded way
 - n Outcome:
 - u Design ofa preliminary set of categories
 - u Recommendation t set up an ISO project based on design as part of the semantic annotation framework project.
- | Summary
 - n Provide more solid foundations for multimdimensionality of DA tag sets
 - n Design consistent truly semantic definitions of core dialogue acts.
 - n Develop agreed definition in the form of iso 12620 and enter in iso registry
 - n Define annotation language with abstract syntax, concrete xml based syntax and

- semantics compliant with laf.
- | Theoretical foundations of da annotation concepts
 - n Relate meaning to dialogue acts
 - n Da has two components for describing utterance meanings:
 - u Semantic content: information which the speakers make available to the addressee
 - u Communication function, capturing the way
- | Multifunctionality
 - n Thank you can be interpreted as
 - u Expression of thanks
 - u Positive feedback
 - u Indication of end of dialogue
- | Annotation must be multidimensional since utterances have multiple meanings
- | Usual information notion of dimension: set of mutually exclusive tags, which is not satisfactory. Request and understanding mutually exclusive?
- | Dimensions in dialogues
 - n Making progress in performing the activity
 - n Providing and eliciting communication
 - n Take and assign turns
 - n Monitor contact, attention, use of time...
 - n Greet, thank, apologize, say goodbye
- | A dimension is an aspect of participating in a dialogue such that
 - n There is a class of dialogue acts
- | Dimensions:
 - n Feedback
 - n Editing
 - n Turn taking
 - n Time
 - n Contact and attention
 - n Opening and closing
 - n Social obligations
- | Meeting adjourned for coffee.
- | Core dimensions and dialogue acts
 - n Data categories from lyrics
 - u Set of 54 core dialogue act types, 24 general purpose functions and 30 dimension specific functions spread over 10 dimensions
 - u Described in iso 12620
- | Validation of lyrics data categories
 - n Usability for human annotations
 - n Inter annotator agreement
 - n 2 trained annotators
 - n Almost perfect agreement ($\kappa \geq 0.80$)
- | Dialogue act annotation language
 - n Da tag components: dimension name, <function name>
 - n <activity, confirm>
 - n <feedback, CheckQuestion>

- n DiaML
 - u Abstract syntax
 - l Conceptual elements to include
 - n Finite set of participants
 - n Finite ordered set of segment begin;end indicators
 - l Concrete syntax
 - u Speaker and addressee
 - u Segments of dialogue behavior
 - u Dimensions
 - u Communicative functions
 - u Optionality: functional dependencies
- l Participants
 - n Geographic considerations

Kiyong: inner group with Kihyong, Harry, and an open group with national members
 Laurent: mature document but need meeting strategy: data category issue (normative vs informative). Format should be left open at this stage.
 Thierry: clarification of what is not to be included in the work.
 Gil: coverage of da. Reported da in news reports.
 Anthropology community should be contacted.
 Laurent questioned the necessity of having a meeting in Moscow.

TDG6: Language Resource Ontologies

Hasida Koiti

- l Issues
 - n Ontologisation
 - n Extension of rdf and ontology framework
 - u Extended rdf instead of xml
 - n Publish trs
 - n Launch iss
- l Ontologisation
 - n Ontology based reformulation
 - u Xml no good lacking standards
 - n Rdf as base description and modeling tool
 - n Ontology as schema
- l Motivations
 - n Dcr model lacks descriptive power
- l Ontologies subsume feature systems

Linguistic Annotation Framework (LAF) (ISO 24612). Nancy Ide described that this standard will be ready for distribution in June. LAF has been coordinated with UMIA. **Graph Annotation Framework (GrAF)** is the instantiation of LAF. Harry Bunt is working on getting discourse acts into LAF, DAMSL, and other annotation frameworks.

Peter Wittenberg says that the Data Category Registry (ISO CAT or ISO 12620) project is still being worked on as part of CLARIN, and will be up later this year. The **ISO CAT and LAF standards are nearly at the end of their allocated time as per ISO**. Lauren Romary is working with ISO management to get more time. Note that data categories from LIRICS have been incorporated into ISO 12620.

According to Harry, not all discourse acts have been included in the categories, (LIRICS has 54), but the list needs to be open for additions. There is disagreement about the number of core discourse acts needed. Study with English and Dutch with two trained annotators working on raw text and audio resulted in **almost perfect interrater reliability**. Additional studies have been done with Italian. Machine learnability investigations are promising. See Harry's presentation or paper for references.

Annotations: "information structures independent of representation format (abstract syntax) Representations: concrete syntax." (Harry). **DiaML** abstract syntax—structure: Speaker, addressee, segment, DiaML-tag. Concrete syntax : define names for all conceptual elements (HTML).

Laurent said that there was a need **for more coordination with the speech community** on dialect acts. Jen asked about coordination with the Anthropology professional organizations. Harry's point of view is that this standard could be used for any culture, but he has **not coordinated with Anthropology organizations**. **It would be good to review the 54 speech acts**. Kiyong said to **contact him to get more materials**. Laurent said that it was difficult to bring in Semitic languages, due to lack of examples and input. He welcomed such input. Jen pointed out that there also needs to be coordination with the speech community with the language codes.

Language Resources Ontologies. Hasida Koiti spoke about the need for "**ontologization**" of all ISO TC 37 standards, with an extension of RDF and ontology framework to more straightforwardly address linguistic information. The proposal is to use extended **RDF** instead of XML and to use **ontology schemas** rather than DTDs, etc. He added that the DCR model lacks this descriptive power (one cannot specify sorts of DCs; cannot specify types of the domain and range of binary relations).

He also pointed out that there are many semantically inconsistent standards (e.g., MPEG-7 > 2000 pages and has many inconsistencies). Note that the W3C recommendation for RDF is at <http://www.w3.org/RDF>. **There are two ISO standards in process: Feature Structure and Feature Structure Documentation**. Features are partial functions. RDF properties are relations in general. Usually feature systems have no taxonomy of features, whereas usual ontologies have taxonomies of properties (e.g., due to `rdfs:subPropertyOf`). A **graph model** is essential. There should be no textual encoding such as XML, although W3C insists on plain text encoding. The ontologies will address FSD; with an extension of RDF. Giles pointed out that specifications are in UML rather than XML.

Proposed standard for References (links) and Citations (text). Daan Broder (sp?) described this project, with particular thanks for the extensive work by Sue Ellen Wright. The practice of establishing references and citations is standardized in W3C via *http URI + # + fragment id* or

by range or by using a service (not standard). Problems with URIs include that: the physical path can be lost with copying, etc; meaningful names may become inappropriate; machine names change; etc. We need a **Persistent Identifier System (PIS)** to separate the resource name from the resource location plus a resolver system to translate names into locations. Current systems that address this problem area include PURL, HS, ARK, XRI, etc. The PIS would avoid link and semantic “rot”. With a PID, one can choose a server when the resource exists on multiple servers. However, this resolving process must be built into applications or made available through plug-ins or http proxy. There is an added layer of resolver administration. Repositories must be able to handle this responsibility in the long term. In Handle (CNRI), every PID is a combination of a pre- and a post- fix. One must find the Local Handle Service, by querying the Global Handle Repository.

There are small sets of tightly related resources (bundles) referenced by links to metadata descriptions than to embed the links (URLs) to the resources. One would replace or augment the URI with a PID. Benefits include being able to better bookmark an archive node or resource.

Laurent asked if we should select certain elements from existing standards and/or if we should standardize more of the identifiers for resources. Daan and Peter Wittenberg said that there were too many resources to provide a single standard. Laurent suggested that the standard be structured with requirements, so that one could say whether a system was compliant with, say, Requirements 1, 4, and 5. Currently, no system is totally compliant. Daan expressed concerns about making requirements too specific (e.g., security). Daan also expressed concern about existing standards for expressing parts of resources. He said that the committee recommended frequent updates to capture these other standards.

Jen asked whether we should have a standard or a Technical Report or standard. Laurent and Peter strongly argued for a standard, particularly for the terminology. Laurent said that such a standard would help in negotiating with other standards committees for interoperability of these standards. Peter and Laurent suggested getting a group together with Sue Ellen and others to determine what should be in the standard and what in annexes or references. Daan said that a standard was important due to extensive investment it would take to implement.

Pre-DIS 24617-1 SemAF/Time. Kiyong Lee reported on work on the draft standard in Hong Kong last year. Items for discussion included a brief introduction of new developments (e.g., semantics of ISO Time-ML). There are seven possible occurrences of “class”. There is a proposal to change this to “eventclass”. There was also a reorganization of the standard. Harry said that there were changes since Hong Kong. He also said that there are some problems (e.g., Time-ML which is currently in a line definition. “Event” includes tag of “polarity”, which he says is more related to the relationship of the event than to the event itself.) He recommended cleaning up Time ML. Nancy said that the Working Group had tried to be agnostic. However, she agreed that it needed some “intellectual cleaning,” such as categorization.

James has finished the comment template. He also agreed to revise the document, working with Kiyong.

Kiyong asked whether we need an abstract syntax at this time. Harry said that it would help to show logical problems. James said that he wished everyone would complete an abstract syntax first in order to have a well-planned standard.

Kiyong discussed that the forward needed to be revised. Note that TimeML refers to the language from that effort. However, ISO TimeML refers to the standard being developed in this document.

The timetable is:

2008-06-05	Kiyong must make a preliminary report to ISO CS
2008-06-30	Revision of Clause 6
2008-07-31	Revision by James Pustejovsky and Kiyong Lee
2008-08-05	Circulation of the draft
2008-09-27/29(PISA)	Final discussion
2008-19-31	Submission of the revised draft to ISO CS for registration as DIS and also for the initiation of the DIS ballot by ISO CS.

Kiyong observed that there is a conflict between another meeting and the ISO TC annual meeting. However, it is important to have the plenary there and to coordinate with the DCR. Laurent suggested that we focus on the DCR structure and content at the international meeting. He recommends having a presentation on the DCR early in the meeting.

Feature Structure Document

ISO-TEI Joint Committee, with Kiyong as project manager.

TC37-4 N245 as a CD 2006:11:30

Completing of WD by 2007-03-31

As a DIS 2007-03-15

As FDIS 2007-08-30

For publication 2008-02-20

Kiyong described the status of the document. Laurent pointed out that TEI can process a new document in six weeks. We should try to go for one document. He said that we need resetting of the target date, creating one new project on feature structures by combining Part 1 FSR and Part 2 FSD, and by re-establishing the liaison with the TEI Consortium. Lou was insisting that the document be completed by the end of September, which Gerald could not do. TEI went ahead and published their guidelines (TEI-5). Note that TEI has moved the first two items together and it is no longer an issue (why?) However, it is difficult to establish communication with TEI. The TI Council is meeting twice per year and is being highly responsive to requests for new features. Laurent observed that we should not try to revise the standard as an ISO standard, or to focus on Part 2 or to bow out. The data is written in XML. Gerard said that he hoped he could finish the draft by the end of July. There is the specification proposer and the decorations. The TEI has corrected typos, accuracy, etc. for Part 1. It is “technically sound even if there may be some scientific problems”.

Kiyong said that the TEI document is a SUBSET of FSD. Laurent pointed out that if we paste the current version of the TEI standard into the FSD, we have the document. There is no formal revision process, since many meetings have gone in many directions. Kiyong agreed to propose this solution to Gerard, who is editing/rewriting the document.

Resolution

To produce a document that would be amenable to integrate the various comments we have received.

Note from Mark: in B17 on 14:40-16:20 on Thursday, there is a poster session on ISO DatCats.

Lexical Markup Framework and LingINFO. Paul Buitelaar is looking at knowledge representation as lexicons. In the LMF Model, this corresponds to "NLP Semantics". He is dealing with homonymy.

LingInfo has lexicalized ontologies, representing terms instead of ontology class labels. (Lexical Semantics is strictly in the (Domain) ontology. There is a lexical ontology enrichment, harvesting ontologies published on the web for question and answering information. There is a LingInfo website with their model.

Multilingual labels for terms in an ontology, which ties into the Ontology Markup Vocabulary (an effort to add metadata to ontologies). You can relate lexical to semantic structures. He is working with WordNet to extract synonyms, deriving translations from Wikipedia, and deriving Morphosyntactic information.

Jen asked how terms were assigned to domains. Paul said he is working in the biomedical domain, where he filters out non-medical usages of terms. Statistical methods could be used. There is also systematic polysemy (e.g., human objects but also acts).

Paul said that he does not understand LMF to know if all of the above could be expressed in LMF or if more development needs to be done on LMF.

Violetta observed that combining lexical and ontological frameworks is much more complicated than is shown in some of the slide examples. Paul debated whether something should be between the semantic and lexical areas. Thierry observed that this connection/integration needs to also occur with the annexes of LMF.

Harmonization. Nancy Ide pointed out that harmonization between ISO standards is needed in the form (representation), content (categories, relationships), and annotation "layers" (types). Nancy earlier had talked about a need to coordinate with OMG. Jen pointed out that Sue Ellen, Monte, and she had had initial meetings. Document consistency is needed in representation of annotations, terminologic use, and standards documents organization.

Alex Fang provided the second part of the presentation, saying that the id of new terminologies needs to be represented in the data registry, identifying variant terminologies. Continued work

will take place at City University of Hong Kong and Vassar, with proposed funding in China and the U.S.

Laurent said that the task is to determine what the core terms are in each area. Nancy said that we could then map these terms.

According to Nancy Ide, the Linguistic Annotation Framework (LAF) (ISO 24612) will be ready for distribution in June. LAF has been coordinated with UMIA. Graph Annotation Framework (GrAF) is the instantiation of LAF. Harry Bunt is working on getting discourse acts into LAF, DAMSL, and other annotation frameworks.

Peter Wittenberg says that the Data Category Registry (ISO CAT or ISO 12620) project is still being worked on as part of CLARIN, and will be up later this year. The ISO CAT and LAF standards are nearly at the end of their allocated time as per ISO. Lauren Romary is working with ISO management to get more time. Note that data categories from LIRICS have been incorporated into ISO 12620.

According to Harry, not all discourse acts have been included in the categories, (LIRICS has 54), but the list needs to be open for additions. There is disagreement about the number of core discourse acts needed. Study with English and Dutch with two trained annotators working on raw text and audio resulted in almost perfect interrater reliability. Additional studies have been done with Italian. Machine learnability investigations are promising. See Harry's presentation or paper for references.

Annotations: "information structures independent of representation format (abstract syntax)
Representations: concrete syntax." (Harry). DiaML abstract syntax—structure: Speaker, addressee, segment, DiaML-tag. Concrete syntax : define names for all conceptual elements (HTML).

Laurent said that there was a need for more coordination with the speech community on dialect acts. Jen asked about coordination with the Anthropology professional organizations. Harry's point of view is that this standard could be used for any culture, but he has not coordinated with Anthropology organizations. It would be good to review the 54 speech acts. Kiyong said to contact him to get more materials. Laurent said that it was difficult to bring in Semitic languages, due to lack of examples and input. He welcomed such input. Jen pointed out that there also needs to be coordination with the speech community with the language codes.

Language Resources Ontologies: Hasidi Koiti spoke about the need for "ontologization" of all ISO TC 37 standards, with an extension of RDF and ontology framework to more straightforwardly address linguistic information. The proposal is to use extended RDF instead of XML and to use ontology schemas rather than DTDs, etc. He added that the DCR model lacks this descriptive power (one cannot specify sorts of DCs; cannot specify types of the domain and range of binary relations).

He also pointed out that there are many semantically inconsistent standards (e.g., MPEG-7>2000 pages and has many inconsistencies). Note that the W3C recommendation for RDF is at <http://www.w3.org/RDF>. There are two ISO standards in process: Feature Structure and Feature Structure Documentation. Features are partial functions. RDF properties are relations in general. Usually feature systems have no taxonomy of features, whereas usual ontologies have taxonomies of properties (e.g., due to `rdfs:subPropertyOf`). A graph model is essential. There should be no textual encoding such as XML, although W3C insists on plain text encoding. The ontologies will address FSD; with an extension of RDF. Giles pointed out that specifications are in UML rather than XML.

Proposed standard for References (links) and Citations (text). Daan Broder (sp?) described this project, with particular thanks for the extensive work by Sue Ellen Wright. The practice of establishing references and citations is standardized in W3C via *http URI + # + fragment id* or by range or by using a service (not standard). Problems with URIs include that: the physical path can be lost with copying, etc; meaningful names may become inappropriate; machine names change; etc. We need a Persistent Identifier System (PIS) to separate the resource name from the resource location plus a resolver system to translate names into locations. Current systems that address this problem area include PURL, HS, ARK, XRI, etc. The PIS would avoid link and semantic “rot”. With a PID, one can choose a server when the resource exists on multiple servers. However, this resolving process must be built into applications or made available through plug-ins or http proxy. There is an added layer of resolver administration. Repositories must be able to handle this responsibility in the long term. In Handle (CNRI), every PID is a combination of a pre- and a post- fix. One must find the Local Handle Service, by querying the Global Handle Repository.

There are small sets of tightly related resources (bundles) referenced by links to metadata descriptions than to embed the links (URLs) to the resources. One would replace or augment the URI with a PID. Benefits include being able to better bookmark an archive node or resource.

Laurent asked if we should select certain elements from existing standards and/or if we should standardize more of the identifiers for resources. Daan and Peter Wittenberg said that there were too many resources to provide a single standard. Laurent suggested that the standard be structured with requirements, so that one could say whether a system was compliant with, say, Requirements 1, 4, and 5. Currently, no system is totally compliant. Daan expressed concerns about making requirements too specific (e.g., security). Daan also expressed concern about existing standards for expressing parts of resources. He said that the committee recommended frequent updates to capture these other standards.

Jen asked whether we should have a standard or a Technical Report or standard. Laurent and Peter strongly argued for a standard, particularly for the terminology. Laurent said that such a standard would help in negotiating with other standards committees for interoperability of these standards. Peter and Laurent suggested getting a group together with Sue Ellen and others to determine what should be in the standard and what in annexes or references. Daan said that a standard was important due to extensive investment it would take to implement.

Pre-DIS 24617-1 SemAF/Time. Kiyong Lee reported on work on the draft standard in Hong Kong last year. Items for discussion included a brief introduction of new developments (e.g., semantics of ISO Time-ML). There are seven possible occurrences of “class”. There is a proposal to change this to “eventclass”. There was also a reorganization of the standard. Harry said that there were changes since Hong Kong. He also said that there are some problems (e.g., Time-ML which is currently in a line definition. “Event” includes tag of “polarity”, which he says is more related to the relationship of the event than to the event itself.) He recommended cleaning up Time ML. Nancy said that the Working Group had tried to be agnostic. However, she agreed that it needed some “intellectual cleaning,” such as categorization.

James has finished the comment template. He also agreed to revise the document, working with Kiyong.

Kiyong asked whether we need an abstract syntax at this time. Harry said that it would help to show logical problems. James said that he wished everyone would complete an abstract syntax first in order to have a well-planned standard.

Kiyong discussed that the forward needed to be revised. Note that TimeML refers to the language from that effort. However, ISO TimeML refers to the standard being developed in this document.

The timetable is:

2008-06-05	Kiyong must make a preliminary report to ISO CS
2008-06-30	Revision of Clause 6
2008-07-31	Revision by James Pustejovsky and Kiyong Lee
2008-08-05	Circulation of the draft
2008-09-27/29(PISA)	Final discussion
2008-19-31	Submission of the revised draft to ISO CS for registration as DIS and also for the initiation of the DIS ballot by ISO CS.

Kiyong observed that there is a conflict between another meeting and the ISO TC annual meeting. However, it is important to have the plenary there and to coordinate with the DCR. Laurent suggested that we focus on the DCR structure and content at the international meeting. He recommends having a presentation on the DCR early in the meeting.

Feature Structure Document

ISO-TEI Joint Committee, with Kiyong as project manager.

TC37-4 N245 as a CD 2006:11:30

Completing of WD by 2007-03-31

As a DIS 2007-03-15

As FDIS 2007-08-30

For publication 2008-02-20

Laurent pointed out that TEI can process a new document in six weeks. We should try to go for one document. He said that we need resetting of the target date, creating one new project on feature structures by combining Part 1 FSR and Part 2 FSD, and by re-establishing the liaison with the TEI Consortium. Lou was insisting that the document be completed by the end of September, which Gerald could not do. TEI went ahead and published their guidelines (TEI-5). Note that TEI has moved the first two items together and it is no longer an issue (why?) However, it is difficult to establish communication with TEI. The TI Council is meeting twice per year and is being highly responsive to requests for new features. Laurent observed that we should not try to revise the standard as an ISO standard, or to focus on Part 2 or to bow out. The data is written in XML. Gerard said that he hoped he could finish the draft by the end of July. There is the specification proposer and the decorations. The TEI has corrected typos, accuracy, etc. for Part 1. It is “technically sound even if there may be some scientific problems”.

Kiyong said that the TEI document is a SUBSET of FSD. Laurent pointed out that if we paste the current version of the TEI standard into the FSD, we have the document. There is no formal revision process, since many meetings have gone in many directions. Kiyong agreed to propose this solution to Gerard, who is editing/rewriting the document.

Resolution

To produce a document that would be amenable to integrate the various comments we have received.

Note from Mark: in B17 on 14:40-16:20 on Thursday, there is a poster session on ISO DatCats.

Lexical Markup Framework and LingINFO. Paul Buitelaar is looking at knowledge representation as lexicons. In the LMF Model, this corresponds to “NLP Semantics”. He is dealing with homonymy.

LingInfo has lexicalized ontologies, representing terms instead of ontology class labels. (Lexical) Semantics is strictly in the (Domain) ontology. There is a lexical ontology enrichment, harvesting ontologies published on the web for question and answering information. There is a LingInfo website with their model.

Multilingual labels for terms in an ontology, which ties into the Ontology Markup Vocabulary (an effort to add metadata to ontologies). You can relate lexical to semantic structures. He is working with WordNet to extract synonyms, deriving translations from Wikipedia, and deriving Morphosyntactic information.

Jen asked how terms were assigned to domains. Paul said he is working in the biomedical domain, where he filters out non-medical usages of terms. Statistical methods could be used. There is also systematic polysemy (e.g., human objects but also acts).

Paul said that he does not understand LMF to know if all of the above could be expressed in LMF or if more development needs to be done on LMF.

Violetta observed that combining lexical and ontological frameworks is **much more complicated** than is shown in some of the slide examples. Paul debated whether something should be between the semantic and lexical areas. Thierry observed that this connection/integration needs to also occur with the annexes of LMF.

Harmonization. Nancy Ide pointed out that harmonization between ISO standards is needed in the form (representation), content (categories, relationships), and annotation “layers” (types). Nancy earlier had talked about a need to coordinate with OMG. Jen pointed out that Sue Ellen, Monte, and she had had initial meetings. Document consistency is needed in representation of annotations, terminologic use, and standards documents organization.

Alex provided the second part of the presentation, saying that the id of new terminologies needs to be represented in the data registry, identifying variant terminologies. Continued work will take place at City University of Hong Kong and Vassar, with proposed funding in China and the U.S.

Laurent said that the task is to determine what the core terms are in each area. Nancy said that we could then map these terms.

Thematic Domain Group 7. Monica Monachini discussed the aim of the new group (based on Honk Kong ISO resolutions with Giles and Nicoletta as chairs). They are trying to show interoperability between lexical, syntactic, and semantic annotation, working with LMF, the DCR, and other standards. LIRICS includes SIMPLE harmonized plurilingual lexicon for 12 languages; KYOTO provides harmonization of synset semantic relations; BootStrep provides an extension for the Biology domain; NEDO provides a definition of harmonized predicate arguman structures and semantic rules.