Towards Interoperability of ISO Standards for Language Resource Management

Kiyong Lee, Laurent Romary

Towards Interoperability of ISO Standards for Language Resource Management*

Kiyong Lee
Korea University
Seoul, Korea
klee@korea.ac.kr

Laurent Romary
Max-Planck Digital Library
Berlin, Germany
laurent.romary@loria.fr

Abstract

This paper reviews and evaluates some of the ISO documents for language resource management. Being developed under ISO/TC 37/SC 4, these documents are at the last stage of approval and publication as international standards. They provide specification languages for the annotation of primary linguistic data and also for the representation of annotated data. The paper makes some technical recommendations to make these specification languages interoperable at the level of both annotation and representation, thus making the use of the resulting language resources sustainable in various areas of language technology applications.

1 Introduction: Purpose and Focus

ISO/TC 37/SC 4 was established in May 2002 to develop ISO documents for language resource management (LRM) that can be accepted as international standards. These documents are designed to provide specification languages for the annotation of primary linguistic data of various types and sources and also for the representation of annotated data in a markup language such as XML. The purpose of this paper is to review some of these ISO documents for their interoperability at the level of both annotation and representation and then to make some specific recommendations for their convergence in a pivotal format for the sustainable use of language resources thus produced. It will, however, focus on the representation schemes of those specification languages for LRM, developed under the Working Group 2 of ISO/TC 37/SC 4.

The paper is organized as follows: section 2 review of some ISO standards for LRM, section 3 Segmentation of primary data, section 4 Identification and reference, section 5 Annotating different layers of linguistic descriptions, section 6 Summary of proposals, and section 7 Concluding remarks.

2 Review of ISO Standards for LRM

ISO/TC 37/SC 4 consists of five working groups (WG's) each with a specific objective. WG 1 aims at working on basic descriptors and mechanisms for language resources, WG 2 on annotation and representation schemes, WG 3 on multilingual text representation, WG 4 on lexical resources, and WG 5 on the workflow of LRM. Except for WG 5, the other four WGs have been activated with a dozen of work items that are now at various stages of development. WG 1 and WG 4 have succeeded in publishing one ISO document each as an international standard: *ISO 24610-1:2006 Feature Structure Representation (FSR)*[1] and *ISO 24613:2008 Lexical Markup Framework (LMF)*.

2.1 Basic Descriptors and Mechanisms

Two of the documents produced by WG 1 lay the basis of developing other LRM-related standards. One is I*SO DIS 24612 Linguistic Annotation Framework*

*A preliminary version of this work was presented by the first author at a research seminar organized by CTL, City University of Hong Kong, 2009-12-10.

[1] Jointly developed with the TEI Consortium and Chapter 5 XML representation of feature structures in *FSR* is contained in *TEI P5*.

(LAF) and another *FSR*. They are basic descriptors and mechanisms for language resource management. *LAF* makes at least the following three proposals for: (1) the adoption of standoff annotation, opposed to inline annotation, (2) a data model of a double-deck structure that clearly demarcates between referential and content structures, and (3) the representation of content structures in feature structures. *FSR* then adopts XML as a markup language and lays out details for the use of XML for representing feature structures. Based on these standards, all of other LRM standards are thus required to follow each of these proposals.

2.2 Representation Schemes

With its aim on LRM representation schemes, WG 2 has been most productive. Four of its documents are at the penultimate stage of approval and publication by ISO as international standards and two other documents are at the stage of committee-internal review. Five new work item proposals have also been made this year alone.

WG documents treat various layers of linguistic descriptions. Two of the documents, *ISO DIS 24611 Morphosyntactic Annotation Framework (MAF)* and *ISO DIS 24615 Syntactic Annotation Framework (SynAF)* treat morphosyntax and syntax, respectively, while *ISO DIS 24614 Word Segmentation of Text (WordSeg)* specifies how a text, say written in Chinese characters, is segmented into words or other segmentation units. The other two are parts of *ISO 24617 Semantic Annotation Framework (SemAF)* on semantic annotation: one treats the annotation of temporal and event-referring expressions in a text with a specification language called *ISO-TimeML* and the other part works on the annotation of dialogue acts and other related pragmatic features in dialogues with *dialML*.[2] WG 2 has also been preparing other parts of *SemAF*: Part 3 named entities, Part 4 space, Part 5 semantic roles, and Part 6 discourse relations.

2.3 Operational Issues for Interoperability

Naively understood, interoperability means operational consistency. When it applies to a system or a set of schemes, each system or set of schemes must not result in internal inconsistencies nor in incompatibility with other systems or other sets of schemes. By interoperability defined operationally as such, some morphosyntactic descriptions as necessitated in one standard (e.g. *WordSeg* or *SemAF*) thus needs to conform to another standard (e.g. *MAF*), which provides such descriptions. These standards are also require to meet the condition of interoperability with other accepted guidelines and recommendations such as TEI guidelines or W3C recommendations. Interoperability is enforced especially when all these layers of linguistic descriptions, each stored in a separate file in standoff manner, converge into one pivotal format to generate one coherent system of information.

Applied to LRM, interoperability operates at three different levels: annotation, representation, and applications. Our work presented here focuses on the interoperability of ISO specifications for LRM at the level of representation schemes.

3 Segmentation of Primary Data

Annotation means adding some notes to some parts of a text or some other types of data. For this purpose, the data is segmented into parts so that some relevant parts, called *markables*, are uniquely identified each with a unique ID and then these parts are referred to in an explicit way.

A text can be segmented into parts of different sizes. *LAF* proposes *base segmentation* that segments primary data into character units. *MAF* introduces two segmentation units, `token` and `wordForm`, while *SynAF* introduces four more grammatical units, `chunk`, `phrase`, `clause`, and `sentence`. *WordSeg* then defines the term word segmentation unit (WSU) as a technical term that may refer to other units than word forms. *SemAF-2* introduces the concept of *functional segment* that allows the segmentation of a single utterance into more than one functional segments each with a distinct communicative function.

3.1 Base Segmentation

In *LAF*, primary data is segmented into characters that are understood as contiguous byte sequences of a specified length. Each annotated data is then associated with a unique base segmentation of primary

[2] This document is at the stage of committee-internal review.

data that defines edges between virtual nodes located between each character in the data. For text, the default is one UTF-8 character. Location indexes are considered to fall between characters, starting at 0. For example, consider the text:

(1) a. Text: `Mia's looked me up.`
b. Base segmentation:

```
0 1 2 3 45 6 7 8 9 0 1 2 3 4 5 6 7 8 9
|M|i|a|'s| |l|o|o|k|e|d| |m|e| |u|p|.|
```

The location indexes for each token are referenced to as spans with start and end.

(2) Location Indexes as Spans:

```
Mia: from="0" to="3"
's: from="3" to="5"
looked: from="6" to="12"
me: from="13" to="15"
up:from="16" to="18"
.: from="18" to="19"
```

3.2 Token and Word Form

In segmenting text or other type of linguistic data into morphosyntactic units, *MAF* introduces two concepts, *token* and *word form*, that are well-known in NLP communities. It then shows ways of representing annotated data in XML with two elements, `<token>` and `<wordForm>`, corresponding to these two concepts.

The tokenization of sample text (1a) given earlier is represented in *MAF*, as below:[3]

(3) Tokenization:

```
<token id="t1" from="0" to="3">
                    Mia</token>
<token id="t2" from="3" to="5">
                    's</token>
<token id="t3" from="6" to="12">
                 looked</token>
<token id="t4" from="13" to="15">
                      me</token>
<token id="t5" from="16" to="18">
                      up</token>
<token id="t6" from"18" to="19">
                      .</token>
```

With this, *MAF* can now show in standoff notation how the sample text is segmented into word forms.

(4) Word Forms:

```
<wordForm lemma= "Mia" tokens="t1"/>
<wordForm lemma= "have" tokens="t2"/>
<wordForm lemma= "look up"
                    tokens="t3 t5"/>
<wordForm lemma= "I" tokens="t4"/>
```

Here two discontiguous tokens `t3` and `t5` are shown to form a single complex word "looked ... up".[4]

3.3 Word Segmentation (WordSeg)

WordSeg consists of two parts: Part 1 treats some basic concepts and general principles of word segmentation of text and Part 2 specific details of word segmentation of Chinese, Japanese and Korean text. This standard is particularly needed to treat written text of Chinese and some other languages in which each character may be treated as a word.

Consider the following fragment of Chinese text:

(5) Chinese Text: 白菜和猪肉

This fragment consists of five characters, each of which can be segmented as words, as a whole meaning "white vegetables accord pig meat". Its normal interpretation, however, is "lettuce goes well with pork". The first two characters "白菜" together are ordinarily treated as a single word that refers to a a particular type of vegetable independent of its color just like the word "blackboard" in English. The last two characters "猪肉" are also understood in the same manner, referring to pork as a one word consisting of two characters.

Such segmentations can be represented in *MAF*. First, each of the five characters is marked up as a token.

(6) Tokenization for Chinese:

```
<token xml:id="t1" from="0" to="1"/>
<token xml:id="t2" from="1" to="2"/>
<token xml:id="t3" from="2" to="3"/>
<token xml:id="t4" from="3" to="4"/>
<token xml:id="t5" from="4" to="5"/>
```

Then we have:

(7) Word Segmentation:

[3] The following representation is in inline notation, but needs to be changed to standoff notation, as is required by *LAF*. See section 2.1.

[4] As will be discussed presently, the attribute names `@id` and `@tokens` need to be replaced by `@xml:id` and `@target`, respectively, in order to be compliant to *TEI P5*.

```
<wordForm lemma="白菜" target="#t1 #t2"/>
<wordForm lemma="和" target="#t3"/>
<wordForm lemma="猪肉" target="#t4 #t5"/>
```

The word form "猪肉" can further be segmented into two word segmentation units, "猪" and "肉". This granularity can be treated by introducing two elements <wfAlt> and <wfColl>, corresponding to <vAlt> and <vColl> in the XML representation of feature structures, and also an attribute @org with a possible value "list" associated with <wfColl>.

(8) Granularity of Word Segmentation

```
<wordForm lemma="白菜" target="#t1#t2"/>
<wordForm lemma="和" target="#t3"/>
<wfAlt>
  <wordForm lemma="猪肉"target="#t4#t5"/>
  <wfColl org="list"/>
    <wordForm lemma="猪" target="#t4"/>
    <wordForm lemma="肉" target="#t5"/>
  </wfColl>
</wfAlt>}
```

3.4 Syntactic Units

ISO DIS 24615 Syntactic Annotation Framework (SynAF) applies to segments of a text that are analyzed as sentences, dealing with their constituency and the dependency among the sentential constituents. In addition to the concept of sentence, *SynAF* introduces the following concepts: word, chunk, phrase, and clause.[5] These concepts need to be incorporated into the syntactic annotation of linguistic data.

3.5 Generalization of Various Types of Segmentation

Various types of segmentation can be generalized with an element <seg> with an attribute @type that has a list of appropriate values that includes characterBased, token, wordForm, word, chunk, phrase, clause, sentence, etc. By this generalization, the element <token>, for instance, is understood as an alternative representation of <seg type="token"/>. This also makes it easier to accommodate any new type of segmentation such as funcSeg for functional segment that is introduced in Part 2 of *SemAF* for dialogue acts.

[5] Applied to non-inflectional languages like Chinese, note that the concepts *word* and *word forms* are identical.

4 Identification and Reference

Base segmentation locates characters in primary data, while tokens are each identified with a unique span in which an associated sequence of characters is located. These tokens with unique id's are then referred to for the unique identification of word forms or larger segments of a text.

4.1 Identifying Segments

In *MAF* and a few other LRM documents, the @id attribute uniquely assigns an ID to a token or other linguistic segments of primary data so that they can be uniquely identified. Then these segments are referred to with their unique ID's for identifying other segments. Here is an example from *MAF*:

(9) Older Version

```
<wordForm id= "w3" tokens="t4 t5"/>
```

This fails to conform to the following conventions in *TEI Guidelines P5*: (a) Prefixing of xml: to @id, (b) Introducing @target which lists the value of each @id attribute that is referred to, and (c) Prefixing of # to each attribute-value that is referred to. These conventions are very simple, but are needed to clearly distinguish what is identified from what is referred to.

MAF can now be easily modified to accommodate these conventions:

(10) Revised:

```
<wordForm xml:id="w3"target="#t4 #t5"/>
Or
<seg type="wordForm"xml:id="w3"
                      target="#t4#t5"/>
```

4.2 Reference in Linking Structures

In addition to the two elements <EVENT> and <TIMEX3>, the following three elements are introduced in *ISO-TimeML* for linking structures: <TLINK>, <SLINK> and <ALINK>. Here is an example for the use of <TLINK>, which anchors an event to a time here:

(11) Mia looked me up yesterday.

(12) Temporal Linking:

```
<EVENT xml:id="e1" pred="LOOK_UP"
 target="#token2 #token4" tense="PAST"/>
<TIMEX3 xml:id="t1" type="DATE"
                    value="2009-12-10"/>
<TLINK target="#e1#t1"relType="DURING"/>
```

Here, the two values `#e1` and `#t1` for the `@target` attribute in `<TLINK>` refer to the value `e1` of the `xml:id` attribute in the `<EVENT>` element and the value `t1` of the `xml:id` attribute in the `<TIMEX3>` element, respectively. Such a specification of the `target` attribute in `<TLINK>` is then understood as relating `e1` to `t1` with the relation type `DURING`. For this representation, *ISO-TimeML* is again modified to be conformant to the *TEI P5*.

5 Annotating Different Layers of Linguistic Descriptions

There are several standards that are being developed under ISO/TC 37/SC 4/WG 2, each of which provides a different representation scheme for annotating a different layer of linguistic descriptions. *MAF* treats morphosyntactic descriptions, *SynAF* syntactic descriptions, each part of *SemAF* a specified subarea of semantic descriptions, while *WordSeg* deals with segmentation issues only. All these linguistic descriptions need to converge on a unified structure of linguistic information. For illustration, we show how *MAF* and *SemAF-Time* provide different types of linguistic content and how they can merge into a single unified structure.

5.1 Morphosyntactic Descriptions (MAF)

After identifying each word form the `<wordForm>` element, *MAF* adds morphosyntactic information content to it in feature structures, as is required of a data model by *LAF*.

(13) Morpho-syntactic Description:

```
<MAF>
  <wordForm xml:id="w1" target="#t1">
    <fs type="morpho-syntax">
      <f name="lemma">
        <string>Mia</string>
      </f>
      <f name="pos">
        <symbol value="noun"/>
      </f>
      <f name="person">
      <symbol value="third"/>
      </f>
      <f name="number">
        <symbol value="singular"/>
      </f>
      <f name="gender">
        <symbol value="feminine"/>
      </f>
    </fs>
  </wordForm>
</MAF>
```

The feature structure represented here in XML is conformant to *FSR* as well as to *TEI P5*. It introduces the element `<symbol>` to specify values of some particular features like grammatical ones that constitute a very restricted set of symbols. The feature name `person`, for instance, has only three values: `first, second`, and `third`.[6]

5.2 Semantic Annotation (SemAF)

SemAF-Time specifies how to annotate temporal and event-related information in a text. Here is an example repeated here:

(14) Primary data:

```
Mia looked me up yesterday.
```

(15) Annotation:

```
<EVENT xml:id="e1" pred="LOOK_UP"
 target="#token2 #token4" pos="VERB"
 tense="PAST"/>
<TIMEX3 xml:id="t1" target="#token5"
 type="DATE" value="2009-11-30"/>
<TLINK target="#e1#t1"relType="DURING"/>
```

This representation conforms to *TEI P5*, but fails to conform to *LAF* at least on two scores. First, it fails to separate referential structures from content structures. Second, it fails to represent content structures in feature structures. Furthermore, it fails to separate morphosyntactic information from semantic information.[7]

5.3 Resolving Interoperability

In order to resolve this incongruence, we first construct the `<MAF>` structure and the `<isoTimeML>` structure separately, with their appropriate name

[6] *Open ANC* does not follow such details of *FSR* on the use of XML for representing feature structures, although it uses *FSR* for representing linguistic annotation content.

[7] The third point has been indicated by Harry Bunt (personal communication).

spaces specified, and then merge them into a single <semAF> structure.

(16) Merging MAF and isoTimeML:

```
<semAF xmlns:iso="http//www.iso.org/semAF>
  <MAF xmlns:iso="http://www.iso.org/maf">
    <wordForm xml:id="w2"
                target="#token2 #token4">
      <fs type="morpho-syntax">
        <f name="lemma">
          <string>look up</string>
        </f>
        <f name="POS">
          <symbol value="VERB"/>
        </f>
      </fs>
    </wordForm>
    <wordForm xml:id="w4"target="#token5">
      <fs  type="morpho-syntax">
        <f name="lemma">
          <string>yesterday</string>
        <f name="POS">
          <symbol value="ADVERB"/>
        </f>
      </fs>
    </wordForm>
  </MAF>
  <isoTimeML xmlns:iso="http://
             www.iso.org/semAF/isoTimeML">
    <EVENT xml:id="e1" target="#w2">
      <fs type="OCCURRENCE">
        <f name="PRED">
          <string>LOOK_UP</string>
        </f>
        <f name="TENSE">
          <symbol value="PAST"/>
        </f>
      </fs>
    </EVENT>
    <TIMEX3 xml:id="t1" target="#w4">
      <fs type="temporal annotation">
        <f name="DATE">
          <string>2009-11-30</string>
        </f>
      </fs>
    </TIMEX3>
    <TLINK xml:id="tL1"target="#e1 #t1">
      <fs type="temporal anchoring">
        <f name="relType">
          <symbol value="DURING"/>
        </f>
      </fs>
    </TLINK>
  </isoTimeML>
</semAF>
```

Suppose the given data is part of a dialogue like the following:

(17) Primary data: Dialogue 1

```
(Gio,fs1): Mia looked me up yesterday.
(Gia,fs2): Did she?
```

SemAF-Dacts can add more information to the annotation given above by inserting the following annotation into the <semAF> structure.[8] Here we have assumed that the dialogue material is identifed as consisting of two utterances u1 and u2 in a pre-processed text segmentation, which constitute functional segments fs1 and fs2, respectively, each with a distinct communicative function.

(18) Dialgoue Annotation:

```
<diaml xmlns="http://www.iso.org/diaml/">
  <maf xmlns="http://www.iso.org/maf/">
  <text xml:id="dialogue1"
      target="#string-range(#text,0,35)">
    <seg type="token" xml:id="token1"
       target="#string-range(#text,0,3)">
    ...
    <seg type="token" xml:id="token9"
     target="#string-range(#text,34,35)">
  </text>
  </maf>
  <text xml:id="dialogue1"
              target="#token1...#token9">
    <seg type="funcSeg" xml:id="fs1"
             target="#token1...#token6"/>
    <seg type="funcSeg" xml:id="fs2"
             target="#token7...#token9"/>
  </text>
  <comment xml:id="dialogue1"
                     target="#fs1 #fs2">
    <fs type="dialogue">
      <f name="participants">
        <vColl org="list">
          <string>Gio</string>
          <string>Gia</string>
        </vColl>
      </f>
    </fs>
  </comment>
  <dialActs xml:id="da1" target="#fs1">
    <fs type="dialogue act">
      <f name="sender">
        <fVal target="#Gio/>
      </f>
```

[8]The annotation given below is slightly modified from the one proposed in Harry et al. (2009) and also from Part 2 of *semAF* for dialogue acts. Furthermore, the two attributes @to and from that specifies a span with start and end in base segmentation are replaced by a sing attribute @target with its values specified with #string-range(#text,i,j), where i stands for start and j for end.

```
      <f name="addressee">
        <fVal target="#Gia"/>
      </f>
      <f name="communicativeFunction">
        <symbol value="inform"/>
    </fs>
  </dialActs>
  <dialActs xml:id="da2" target="#fs2">
    <fs type="dialogue act">
      <f name="sender">
        <fVal target="#Gia/> </f>
      <f name="addressee">
        <fVal target="#Gio"/> </f>
      <f name="communicativeFunction">
        <symbol value="checkQuestion"/>
    </fs>
  </dialActs>
</dialml>
```

6 Proposals for Representational Interoperability

A list of proposals is given here to show how some of the key requirements of *TEI*, *LAF*, and *FSR* can be met for the interoperability of representing lingustic annotations. First, we discuss a set of TEI requirements that are laid out in *TEI P5* for segmentation, identification and reference. Then we make proposals following the three key requirements laid out by *LAF*: (1) Standoff annotation opposed to inline annotation, (2) demarcation of referencing and content structures, and (3) representation of content structures in feature structures. These requirements need to be met for the operation of a pivotal format that allows the merging, extension, and exchange of different types of annotation.

6.1 Segmentation of Primary Data

For the segmentation of text or any data in general, *TEI* introduces the `<seg>` element with a list of appropriate types. We list these types as follows:

(19) **Proposal 1:** Introduce the `<seg>` element, with the following `@type` attribute and with a list of appropriate vavlues `baseSeg`, `token`, `wordForm`, `chunk`, `phrase`, `clause`, `sent` and `funcSeg`:

(20) Segmentation:

```
a. <seg type="token"/>
b. <seg type="wordForm"/>
c. <seg type="funcSeg"/>
```

We may, however, adopt an alternative representation as allowed in *TEI*.

(21) **Convention 1:** Interpret `<token>`, `<wordForm>`, etc. as alternatives (or syntactic sugar) to: `<seg type="token">`, `<seg type="wordForm">`, etc., respectively.

This convention allows elements such as `<token>`, `<wordForm>`, etc. corresponding to the values of the `@type` attribute, as listed in Proposal 1. The `<token>` element is then understood as standing for `<seg type="token">`.

6.2 Identication and Reference

Following *TEI* again, we adopt the following:

(22) **Proposal 2**
a. The `@xml:id` should be systematically used and references to element identifier made through URIs.
b. Introduce the `@target` attribute with a sequence of URIs as value.

(23) Illustration for Prop 2:

```
<seg type="wordForm" xml:id="word2"
 target="#token2 #token4"/>
```

The reference mechanism may extend to linking structures, as illustrated below:

(24) Reference for Linking:

```
<TLINK type="temporal" target="#e1 #t1">
  <fs type="temporal anchoring">
    <f name="relType">
      <symbol value="DURING"/>
    </f>
  </fs>
</TLINK>
```

Here the `<TLINK>` is understood as `<link type="temporal">`.

6.3 Standoff Annotation

Now following *LAF*, we propose the following:

Proposal 3: Adopt standoff annotation.

The following annotation exemplifies how one should adopt a syntax that is continuous to the URI reference. It adopts the `string-range()` `XPointer scheme` (with the hypothesis that an

xml:base has been defined in the parent element):[9]

(25) Illustration for Prop 3:

a. Inline:

```
<seg type="token" xml:id="token1"
target="#string-range(#text,0,3)">
                          Mia</seg>
```

b. Standoff:

```
<seg type="token" xml:id="token1"
 target="#string-range(#text,0,3)"
 form="Mia"/>
```

6.4 Data Model: Referential Structure and Annotation Content Structure

Being conformant to *LAF*, we prose the following:

(26) **Proposal 4:**

a. Construct each data model as a double-deck structure, consisting of (1) referential structure and (2) annotation content structure.

b. Specify the attributes @xml:id and target with appropriate values in each referential structure.

c. Represent each annotation content structure in feature structures.

(27) Illustration for Prop 4:

```
a. Mia looked me up.
b. <LRM:annLRM
     xmlns="http://www.iso.org/LRM
     xmlns="http://www.tei-c.org/ns/1.0">
   <MAF:MAF
     xmlns:MAF="http://www.iso.org/maf>
     <MAF:wordForm xml:id="w2"
              target="#token2 #token4">
       <fs type="morpho-syntax">
         <f name="LEMMA">
           <string>look up</strng>
         </f>
         <f name="POS">
           <symbol value="VERB"/>
         </f>
       </wordForm>
   </MAF:MAF>
   <TimeML:isoTimeML xmlns:TimeML=
   "http://www.iso.org/semAF/isoTimeML">
     <TimeML:EVENT xml:id="e1"
                          target="#w2">
       <fs type="OCCURRENCE">
         <f name="PRED">
           <string>LOOK_UP</string>
         </f>
         <f name="tense">
           <symbol value="PAST"/>
         </f>
       </fs>
     </TimeML:EVENT>
   </TimeML:isoTimeML>
   </LRM:annLRM>
```

Each of the elemnts <annLRM>, <MAF> and isoTimeML provide their respective namespaces, while the elements <wordForm> and <EVENT> each represent their respective referential structure. To each of these referential structures a feature structure is embedded.

7 Concluding Remarks

For ISO, standards are documents. Applicable to the standardization of LRM, the number of the documents has amounted to more than a dozen, including six newly proposed work items. For the interoperability of these standards, their representation schemes have been examined especially with respect to some of the requirements laid out by *TEI*, *LAF*, and *FSR*. All the issues discussed are technical in the sense that their resolutions are basically constrained by the very conventional nature of the representation language, namely XML, adopted for the annotation of language resources. It is hoped in further study that our proposals be elaborated with details and incorporated into one of the ISO documents that deals with basic representational requirements.

Acknowledgment

The first author wishes to thank Prof. Jonathan Webster, Dr. Alex Fang, and other faculty and staff members of Department of Chinese, Translation and Linguistics, City University of Hong Kong, for their invitation that enabled him to stay at this department as visiting professor and to complete the drafting of the paper and also to all the members of the Dialogue Systems Group of City University for their support and comments.

References

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, Claudia Soria,

[9] Briefly discussed in section 5.3, footnote 8.

David Traum, Kiyong Lee, and Laurent Romary. 2009. Towards an ISO standard for dialogue act annotation. Unpublished.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Journal of Naturaql Language Engineering*, 10:3-4, 211-225.

Nancy Ide and Laurent Romary. 2007. Towards international standards for language resources. In Dybkjaer, L., Hemsen, H., Minker, W. (Eds.), Evaluation of Text and Speech Systems. Springer, 263-84.

Kiyong Lee, Lou Burnard, Laurent Romary, Eric de la Clergerie, Thierry Declerck, Syd Bauman, Harry Bunt, Lionel Clement, Tomaz Erjavec, Azim Roussanaly, and Claude Roux. 2004. Towards an international standard on feature structure representation. Workshop Proceedings of The Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon.

Kiyong Lee, James Pustejovsky, and Branimir Boguraev. 2006. Towards an international standard for annotating temporal information. *The Third International Conference on Terminology, Standardization and Technology Transfer*, Beijing.

International Organization for Standardization. 2006. *ISO 24610-1:2006 Language Resource Management - Feature Structures - Part 1: Feature Structure Representation (FSR)*. ISO/TC 37/SC 4/WG 1 and TEI Consortium.

International Organization for Standardization (ISO). 2008. *ISO 24613:2008 Language Resource Management - Lexical Markup Framework (LMF)*. ISO/TC 37/SC 4/WG 4.

International Organization for Standardization (ISO). 2008. *ISO DIS 24611 Language Resource Management - Morpho-syntactic Annotation Framework (MAF)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO DIS 24614-1 Language Resource Management - Word Segmentation of Text - Part 1: Basic Concepts and General Principles (WordSeg-1)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO CD 24614-2 Language Resource Management - Word Segmentation of Text - Part 2: Chinese, Japanese, and Korean (WordSeg-2)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO DIS 24615 Language Resource Management - Syntactic Annotation Framework (SynAF)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO CD 24616 Language Resource Management - Multilingual Information Framework (MLIF)*. ISO/TC 37/SC 4/WG 3.

International Organization for Standardization (ISO). 2009. *ISO DIS 24617-1 Language Resource Management - Semantic Annotation Framework - Part 1: Time and Events (SemAF-Time)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO CD 24617-2 Language Resource Management - Semantic Annotation Framework - Part 2: Dialgoue Acts (SemAF-Dacts)*. ISO/TC 37/SC 4/WG 2.

International Organization of Standardization. 2009. *ISO PWI 24617-3 Language Resource Management - Semantic Annotation Framework - Part 3: Named Entities (SemAF-NE)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO PWI 24617-4 Language Resource Management - Semantic Annotation Framework - Part 4: Space (ISO-Space)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO PWI 24617-5 Language Resource Management - Semantic Annotation Framework - Part 5: Semantic Roles (SemAF-semRoles)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO PWI 24617-6 Language Resource Management - Semantic Annotation Framework - Part 6: Discourse Relations (SemAF-DRels)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization (ISO). 2009. *ISO DIS 24619 Language Resource Management - Persistent Identification and Access for Language Technology Applications (PID)*. ISO/TC 37/SC 4/WG 1.

International Organization for Standardization (ISO). 2009. *ISO PWI 24620-1 Simplified Natural Language - Part 1: Basic Concepts and General Principles (simppL-1)*. ISO/TC 37/SC 4/WG X and Object Management Group (OMG).

International Organization for Standardization (ISO). 2009. *ISO PWI 24621 Language Resource Management - Date-Timne Vocabulary (tempVoc)*. ISO/TC 37/SC 4/WG 2 and Object Management Group (OMG).

Interjeet Mani, James Pustejovsky, and Rob Gaizauskas (eds.). 2005. *The Language of Time: A Reader*. Oxford University Press, Oxford.

James Pustejovsky, Robert Ingria, Roser Sauri, Jose Gastano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Christopher Habel. 2005. The specification language TimeML. in Mani et al.(eds.), 2005.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2009. *ISO-TimeML: an International Standard for Semantic Annotation*. Unpublished.