Fifteen (and a half) Years of the TEI

A Retrospective and Prospective

Nancy Ide

Department of Computer Science

Vassar College

Brief History of the TEI and me

- 1987:
  - Obtained NEH grant to hold a meetingof experts on text encoding standards
  - Convened the Poughkeepsie meeting that led to creation of the TEI
- First Steering Committee chair
- Steering Committee until 1996(?)
- Dictionary committee - drafted bulk of the specs

What this talk is about

- Some thoughts on why certain communities -- especially the computational linguistics community -- have not used the TEI
  - Use this as a basis for considering what could be done in the TEI
  - CL community is a heavy user of markup/annotation

More History

- The Association for Computational Linguistics was one of the three organizations supporting and governing the TEI
- But the ACL community never adopted the TEI, even when they began using SGML/XML for corpus annotation

Ca. 1990

- TEI representatives (N. Ide, A. Zampolli, D. Walker, S. Hockey, M. Sperberg-McQueen) met with representatives of the US National Science Foundation and DARPA to discuss possibility of funding for the TEI
- Mark Liberman, head of the Linguistic Data Consortium and informal advisor to funding agencies, also in attendance

The Put Down

- Liberman placed a stack of emails on the table, responses from members of the CL community he had polled to see if they felt any need for development of encoding standards
- Summary:
  - CL does not need standards
  - Just use any format and transduce to another as needed

Was He Right?

Yes…and No

Why yes?

A single standard is not appropriate for CL use/applications

Why no?

One encoding scheme does not necessarily map to another

More on these points later….

Background

- What do computational linguists need to encode?
  - Corpora of written and spoken data
    - Written data: markup for gross logical structure (chapter, headings, titles, section, footnote, etc. down to the level of paragraph)
    - Spoken data: turn and utterance, timestamps
  - Linguistic annotation of these data
    - Part of speech, syntactic analysis, co-reference, discourse structure, named entities…
    - Speech data: prosody, phonology, etc.
  - Alignment of parallel data
    - Parallel translations of the same text
    - Alignment of speech signal, orthographic transcription, etc.
  - Computational lexicons, term banks, etc.

Special Problems

- May encode any/all of these annotations, possibly in incremental steps or at different sites
  - Too much information in a single document
  - Problem of overlapping hierarchies
  - Problem of coordinating work on same documents if done at different sites, different times
  - Also need for incremental processing, separability of annotations during processing

Norm in the early '90's

- Most sites had in-house software with special or proprietary formats for data that their software processes
- Not much exchange going on in early '90's
  - But this of course changed as the costs of corpus annotation became large

Aside…

- Here is where CL first saw that Liberman was wrong
  - Transducing one format to another is not just a matter of replacing one representation for another!
    - If the underlying data model is different, may be no possible mapping
    - Very often information about the structure and content of the data is hidden in the processor's code--not retrievable

Simple example:

(VERB-SUBJ ((DET-POSS) (N-N-MOD)))

Have to know the meaning of VERB-SUBJ

List or set of alternatives?

The Real Problems

- Apart from a strong case of the "not invented here" syndrome, there are several valid reasons why the TEI Guidelines were never adopted by the CL community, and many science-oriented communities in general

The TEI Guidelines are too extensive

- Cover a very wide range of document types and phenomena

Hard to find only what you need

- Offer solutions for encoding a great variety of textual facts, but do not recommend which facts are to be encoded in a document treating a specific sub-area

Hard to know only what you need

The TEI Guidelines are too general

- Intended to be maximally applicable across a wide range of disciplines

Therefore:

- Often take encoding solutions to the highest possible level of abstraction
- Allow multiple different ways to encode the same phenomenon

Inhibits validation

- Tension between the generality of an encoding scheme and the ability to validate
  - **Over-generative DTDs** allow tag sequences which, for any given text, may not be valid
    - Biggest culprit is "paragraph/phrase content allowed everywhere"
- Tight validation extremely important for CL when creating/annotating large corpora
  - Check encoding consistency

Other Enemies of Validation

- **Use of abstract, general tags**
  - E.g., use of a general tag such as to mark different hierarchical divisions of a text disallows constraints on what can appear within a given text division
    - Impossible to ensure that tighter structural constraints for a given book are observed (e.g., titles do not appear within chapters, or paragraph does not appear outside the chapter level, etc.)
- **Multiple ways to encode the same phenomenon**
  - No assurance of same encoding from person to person, corpus to corpus

The TEI Guidelines are too detailed

- Often provide highly detailed, esoteric object descriptions
- E.g., <persname>

```
<persName>
   <title>Sir</title>
   <foreName>Edward</foreName>
   <surName type="linked">Bulwer-Lytton</surName>
   <rolename>Barron Lytton of
      <placeName>Kenworth</placeName>
   </roleName>
</persName>
```

Used only in very specific applications

Markup Semantics are Informal

- General suggestions, but rely on user to apply a given tag "as appropriate"
- E.g, **<w>** tag may appear in a legal syntactic context in an interchanged text, but sender and receiver may not have the same understanding of the semantics

- For computational linguistics, can have big impact
  - impairs immediate reusability
  - E.g., a simple word count or the content of a lexicon created from the text could vary considerably depending on the definition
- Re-use of same tag for different contexts is over-general
  - E.g., <name> in paragraph content very different from <name> in the header
    - Not doing linguistic analysis of names in header, so no need for complex internal structure for the tag in this context

The TEI Guidelines are not written for the CL community

- Style of presentation is not at all what scientists are used to
  - Describe everything from start to finish
    - No easy top-down summary at front
  - Prose descriptions (sometimes lengthy and dense) vs. enumerations, bulleted lists, etc.
  - No rationale provided for choices made

The TEI Guidelines do not provide needed tags for CL

- No tags for specific categories of information needed for CL
  - Morpho-syntactic analyses are the first obvious thing lacking
- Possible to do what is needed, but in most cases must use generic tags
  - E.g., <seg>
- …or a possibly far too bulky mechanism like feature structures
  - A bit too much when you have simple information and no need for operations over feature structures

The TEI DTD requires everything in One Big Document

- CL oriented toward heavy processing of the data
  - Huge number of potentially unused tags increases overhead
  - Often need to use only one logical part of the encoding (specific annotation type etc.)
  - Annotation itself often performed automatically and incrementally
- Documents with multiple types of annotation can quickly become unwieldy or even invalid
  - Overlapping hierarchy problem

An Attempted Solution

- In 1994, developed the Corpus Encoding Standard (CES, now XCES) to answer all these problems
  - Very reduced subset of TEI tags
  - Tag content drastically restricted
  - Precise guidance on what to encode and how
  - More precise tag semantics

- **"Stand-off" markup introduced**
  - Allowed for separation of annotation and primary text document
  - Annotation docs linked to primary, forming a hyper-linked "document"
  - Separate DTD and precise tags for morpho-syntactic annotation
  - Separate DTD for links between/among parallel data
  - Overcame overlapping hierarchy problem
  - Allowed multiple annotations of same type for a given document

    http://www.cs.vassar.edu/CES

    http://www.xml-ces.org
  - Used for the American National Corpus

    http://AmericanNationalCorpus.org

The Point

- Highly streamlined tag sets are more usable than one large one
- Tighter restrictions on tag content (possibly in varying contexts) is desirable for validation
- Ability to link multiple documents and regard as one "hyper-document", or separate out certain annotations, is useful

Recommendation

- Along the "Pizza Chef" model, TEI should provide means for users to custom-build tag sets and specify their content
- Automatically generate the XML schema
- Automatically generate relevant documentation

So…was Liberman right or wrong?

- He was probably **right** about not needing to develop specific tag sets, or even use a specific framework (e.g. XML, LISP, etc.)
- He was **right** that the ideal is to allow everyone to use own encoding scheme and transduce to other formats
- *But he was **wrong** that transduction is possible (easy?) without any standards*

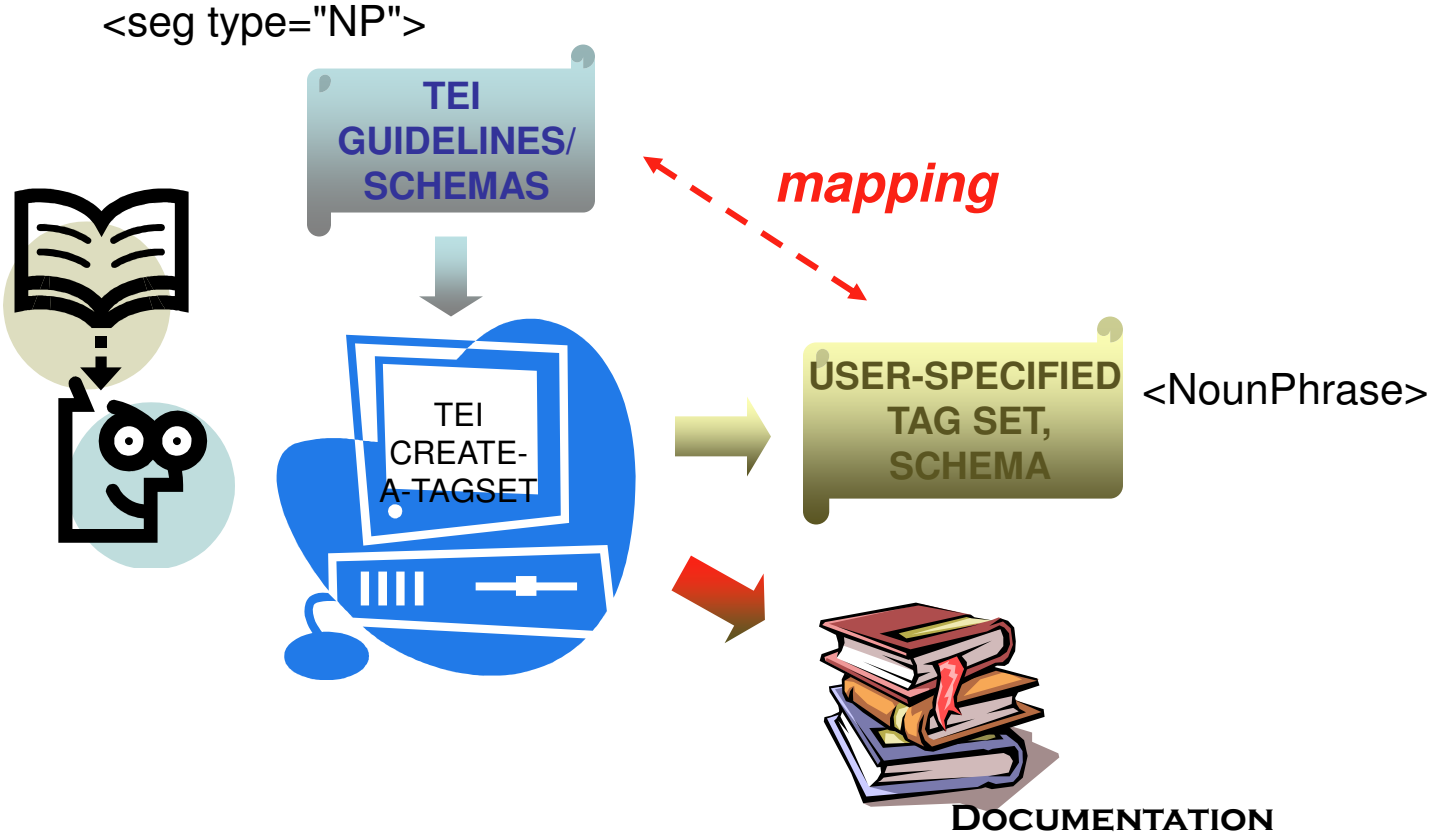
But it is the underlying DATA MODELS that need to be standardized, not the tags that instantiate them

Should the TEI go out of business?

- Knowingly or unknowingly, the TEI has been in the data modeling business since it began…
  - E.g., <persName> : detailed analysis is a *model* of the components of a name
- …so the answer is "no".
- But perhaps a slightly different perspective on the work, wherein there is less focus on defining specific tags, and more on the model behind

(Yet Another Recommendation)

Allow the option for user-defined tag names that map (via XML schema equivalences) to TEI tags



What Are the CL Folks Doing Now?

- Working on generalized standards
  - **ISO TC37 SC4** : **Language Resources**
  - Developing a *Linguistic Annotation Framework* for encoding linguistic resources
    - Separation of
      - User-defined encoding scheme
      - Abstract, generic "pivot" format for interchange
        - ✓ Map between the two

Key is assuring the data model behind each is the same

ISO TC37 SC4

- Developing a registry of **data categories** for linguistic info
  - Agreed upon categories, or variants of different categories
  - Annotators can refer to categories in an encoding via a URI
  - If deviating from a defined category, provide a formal description of differences

Ways the TEI Can Go Forward

- Exploit ideas and expertise in SC4 concerning
  - data models
  - mappings from concrete to abstract syntax, etc.
- Feed into SC4
  - Already:
    - Feature structure encoding scheme
    - Stand-off Working Group has been involved with LAF development
  - NO need to abandon the TEI or melt it into ISO, but simply ensure consistency, avoid duplication of effort, and gain from each others' experience and expertise

Create a synergy between the groups

Conclusion

- The TEI was and is a good thing
- Needs to move forward with the technology and advances in thinking
  - XML, XML Schemas
    - Enable a lot of things not yet exploited by TEI
    - Easier to move toward **modularity**
  - RDF, RDF Schemas, OWL…?
    - Not sure if this works in TEI, but should consider that many users may turn to these schemes in the future to answer encoding needs in web data

- Needs to form partnerships with groups like ISO TC37 SC4 to ensure that the TEI is at least consistent withtheir work
  - Better yet, taken into account in the process of their work
  - Even better yet, integrated (to one level or another) with their work
- May need a shift in perspective
  - Away from specific tag sets, more focus on data modeling

Thank you

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

Chercheur Associé
LORIA/INRIA
Vandeouvre-lès-Nancy, France