Journal of the Text Encoding Initiative

Issue 3 (November 2012)
TEI and Linguistics

Maik Stührenberg

The TEI and Current Standards for Structuring Linguistic Data

An Overview

Electronic reference
Maik Stührenberg, « The TEI and Current Standards for Structuring Linguistic Data », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 15 October 2012, connection on 05 November 2012.
URL : http://jtei.revues.org/523 ; DOI : 10.4000/jtei.523

Publisher: Text Encoding Initiative Consortium
http://jtei.revues.org
http://www.revues.org

Document available online on:
http://jtei.revues.org/523
Document automatically generated on 05 November 2012.

Maik Stührenberg

The TEI and Current Standards for Structuring Linguistic Data

An Overview

1. Introduction

1 During the last decade linguistic annotation of corpora has undergone a substantial change. While in the late 20th century annotation formats were developed and used exclusively for projects or within small communities, we now have a large number of standardization efforts carried out by the International Organization for Standardization (ISO), addressing, in particular, new advancements in technology such as very large and multiply annotated corpora. An overview is given by Ide and Romary (2007) and Declerck et al. (2007).

2 In addition, these standardization efforts are increasingly adopted in international projects such as CLARIN (Common Language Resources and Technology Infrastructure) and FLARENET (Fostering Language Resources Network).[1] Both projects involve harmonization of formats and standards for language resources and technology with the goal of making these much more accessible to researchers via component metadata registries (see Broeder et al. 2011) and by providing guidelines to choose particular specifications (see Monachini et al. 2011).

3 Of course, international standards are not developed in isolation, without any reference to established de facto standards such as the TEI Guidelines. However, there are some differences that can be observed when comparing the TEI Guidelines to these specifications with respect to various aspects of markup languages such as the formal model, the notation, and the annotation model.

4 After a short overview of the process of standardization of international standards, we will contrast this process with the development of community-based specifications, such as the TEI Guidelines. After this introduction, a number of ISO standards that deal with the annotation of language corpora will be examined. The TEI's influence on the development of these standards will then be discussed.

This paper will conclude with recommendations for scholars and researchers that deal with linguistically annotated corpora.

2. Current International Standards

2.1. International Standardization

5 The term *standard* can have two meanings. On the one hand, the term can depict international (or national) industry norms and standards—that is, specifications developed by organizations that have been assigned to this task, such as ANSI (American National Standards Institute) in the USA or DIN (Deutsches Institut für Normung) in Germany. Such standards are called *de jure standards*.

On the other hand, there are also *de facto* (or *market-driven*) *standards*, i.e., specifications that are not endorsed by a standards organization but have achieved a greater popularity compared to similar specifications. An obvious example of such a de facto standard is the original file format of Microsoft Word: the ubiquitous "doc" format. In this case, the status of the specification is based on the dominant market position of the respective company. Another example is the tagset of the TEI Guidelines, the status of which can be explained by its broad acceptance by scholars around the world.

6 De jure standards are developed by international committees, usually under the auspices of the International Organization for Standardization (ISO) and comprising members from various national standards bodies. ISO, for example, has technical committees (TC), divided into subcommittees (SC) and then into working groups (WG) chartered to work on a distinctive topic. But the work of developing a standard often begins in one or more national bodies, since technical committees are made up of national representatives of various stakeholders

such as industry, NGO, government or academia. Therefore, each national organization for standardization (a member body) decides to participate in a number of technical committees. These national bodies often reflect the structure of ISO, allowing for straightforward collaboration between corresponding committees in different countries.

7 A relevant ISO subcommittee in the field of linguistic annotation is ISO/TC 37/SC 4 (in this case, "SC" is for subcommittee 4) called "Language Resource Management", of the technical committee "Terminology and other Language and Content Resources". It is divided into six working groups (WG):

- WG 1: Basic descriptors and mechanisms for language resources
- WG 2: Annotation and representation schemes
- WG 3: Multilingual information representation
- WG 4: Lexical resources
- WG 5: Workflow of language resource management
- WG 6: Linguistic annotation.[2]

8

These working groups develop relevant specifications for the field of linguistic annotation.

9 ISO has a protocol for the proposal process (International Organization for Standardization/ International Electrotechnical Commission 2012) in which proposals must pass through seven stages, each of which takes some time, before becoming official standards:

- Preliminary stage
- Proposal stage
- Preparatory stage
- Committee stage
- Enquiry stage
- Approval stage
- Publication stage

10 The first stage marks the introduction of a Preliminary Work Item (PWI), which can be introduced by members of the working group or by outside interested parties. After a positive internal review, it becomes a New Work Item Proposal (NP). At that time it reaches the proposal stage, in which the so-called P-members ("participating members") of the respective committee (or sub-committee) have to vote in favor or against the further pursuit of this item.[3] If the majority of the P-members cast a positive vote and at least five P-members signal a willingness to participate in the standardization process, the NP is added as a new project of the WG, reaching the beginning of the preparatory stage.

11 In each of the following stages the status of the proposal changes according to substantial improvements that have been made. The committee stage is the first stage at which the Committee Draft (CD), as it's then called, is commented on by national bodies of the TC/SC. This stage ends when all technical issues have been resolved. In that case the CD is transformed into a Draft International Standard (DIS) and enters the enquiry stage.

12

At this stage the DIS will be circulated to all national bodies for a ballot. A vote can be either positive, negative, or an abstention; in the two former cases the vote may be accompanied by editorial or technical comments. The DIS is approved if a two-thirds majority of the P-members' votes are in favor and not more than one-quarter of the total votes cast are negative. In that case it will be registered as a Final Draft International Standard (FDIS), proceeding to the approval stage.[4]

13 From this point onwards the text of the FDIS is usually not publicly available for free (although there are exceptions to this rule). As a result, researchers often consult and cite Committee Drafts or Draft International Standards in their work. However, such a time-consuming and consensus-driven process means that major changes often exist between draft versions and the final International Standard. In contrast, openly developed standards such as the TEI Guidelines are often publicly available both as drafts and final versions, which eases the adoption of changes between different versions.

14 The boundaries between de facto and de jure standards can be very weak; in fact, sometimes de facto standards became de jure standards. For example, Simons (2007) explains the long process of developing a standard for describing language codes, starting from Ethnologue and ending with the International Standard ISO 639-3:2007.[5]

15 In the next section we will discuss some de jure standards that have been developed in ISO/TC 37/SC 4 that may affect the work of current and future linguists.[6]

2.2. Feature Structures (FS)

16 Feature Structures are general-purpose data structures consisting of a named feature and its value (or values). Complex feature structures contain a group of individual features allowing for a representation of various kinds of information.

In linguistics, feature structures are best known as part of Head-driven Phrase Structure Grammar (HPSG).[7]

17 Feature structure representations have been a part of the TEI Guidelines from the very beginning.[8]

However, during the transition from P4 to P5 a substantial amount of work was undertaken to improve the tag set and to clarify its underlying formal logic.

18 The following is an example of a TEI-based linguistic feature structure:

```
<fs>
 <f name="CAT">
  <symbol value="np" />
 </f>
 <f name="AGR">
  <fs>
   <f name="NUM">
    <symbol value="sing" />
   </f>
   <f name="PER" />
    <symbol value="third" />
   </f>
  </fs>
 </f>
</fs>
```

Figure 1: TEI-based feature structure for a linguistic annotation (from Stegmann and Witt 2009).

19 This feature structure consists of two features. The first, named “CAT”, is a simple feature that has the atomic feature value “np”. The second, named “AGR” is a complex feature (that is, its value consists of other feature structures), containing the features “NUM” and “PER”.

20 A few key players in the TEI community submitted the P5 revision of the feature structure annotation format for standardization as the two-part ISO standard 24610. While the first part, ISO 24610-1:2006, describes feature structures (including the representation format shown in the example above and an informal overview of the basic characteristics of feature structures), the second part, ISO 24610-2:2011, discusses feature system declaration described in Chapter 18.11 of the TEI Guidelines.

21 Both parts of ISO 24610 use a RELAX NG grammar that is a subset of the TEI’s P5 document grammar with only slight changes (for example, a different root element). As one may observe, there is a five-year gap between the two parts of ISO 24610. In addition, ISO 24610-1 was scheduled for a regular revision that should have been finished in early 2012. However, due to time constraints on the part of the involved experts, work on the Committee Draft for the revision has been put on hold, leaving ISO 24610-1:2006 as the current version.

2.3. The Linguistic Annotation Framework (LAF)

22 Development of the Linguistic Annotation Framework began in 2005, and it became an approved standard in 2012 (ISO 24612). Its goal is to establish a definitive standard based on widely used de facto standards such as the TEI, the Corpus Encoding Standard (CES, see Ide 1998), and its successor XCES (Ide et al. 2000).

23 LAF provides a framework for representing linguistic annotation of various kinds. It includes an abstract data model for general-purpose linguistic annotation (in contrast to more specific

annotation formats such as the Morpho-Syntactic Annotation Framework discussed in the next section) and an XML serialization format called Graph Annotation Format (GrAF), which serves as a pivot format for mapping between user-defined annotation formats. The data model consists of three parts: (1) anchors that define regions by referencing locations in the primary data (that is, the data to be annotated); (2) a graph structure, consisting of nodes, edges and links to the before-mentioned regions; and (3) an annotation structure comprising a directed graph referencing regions or other annotations. The nodes in this graph are associated with feature structures providing the annotation content. LAF does not include data categories but instead relies on ISO 12620:2009, the International Standard for describing data categories, and on ISOcat, an implementation of ISO 12620:2009 developed in ISO/TC 37/SC 3.[9]

24 A language resource conforming to LAF consists of the primary data; a base segmentation (that is, at least one document that provides anchors and therefore defines regions of the primary data); a number of annotation documents containing nodes, edges and feature structures; and a set of header files (metadata). By storing primary data and annotation in separate files, LAF uses stand-off annotation (see Thompson and McKelvie 1997), similar to CES and XCES, to more easily encode overlapping and discontiguous regions than if these were encoded in a single file. The anchors are nodes that are located between base units of the primary data. Depending on the type of primary data (text, audio, video, or other) the base unit can be a character, a segment of time, or another useful unit of segmentation. An annotation document contains annotations associated with the nodes in the graph that reference regions of the primary data. While stand-off annotation would allow the combination of several linguistic annotation layers into a single annotation document (see Stührenberg and Jettka 2009), the standard recommends the use of separate annotation files for the purpose of exchange.

25 Figure 2 shows a fragment of an example annotation document containing both a header, nodes, edges and annotations (taken from ISO/FDIS 24612).

```
<?xml version="1.0" encoding="UTF-8"?>
<graph xmlns="http://www.xces.org/ns/GrAF/1.0/">
 <graphHeader>
  <labelsDecl>
   <labelUsage label="fullTextAnnotation" occurs="1"/>
   <labelUsage label="Target" occurs="171"/>
   <labelUsage label="FE" occurs="372"/>
   <labelUsage label="sentence" occurs="32"/>
   <labelUsage label="annotationSet" occurs="171"/>
   <labelUsage label="NamedEntity" occurs="32"/>
  </labelsDecl>
  <dependencies>
   <dependsOn type="fntok"/>
  </dependencies>
  <annotationSpaces>
   <annotationSpace as. default="true"/>
  </annotationSpaces>
 </graphHeader>
 <node xml:/>
 <a label="FE" ref="fn-n156">
  <fs>
   <f name="FE" value="Speaker"/>
   <f name="rank" value="1"/>
   <f name="GF" value="Ext"/>
   <f name="PT" value="NP"/>
  </fs>
 </a>
 <!-- [...] -->
 <edge xml: from="fn-n156" to="fn-n133"/>
 <!-- [...] -->
 <region xml: anchors="980 9190"/>
 <region xml: anchors="980 993"/>
 <!-- [...] -->
 <node xml:>
  <link targets="r1"/>
 </node>
 <node xml:>
```

```
  <link targets="r2"/>
 </node>
 <!-- [...] -->
 <a label="R Gesture Units 1" ref="a232"/>
 <a label="preparation" ref="a233"/>
</graph>
```

Figure 2: An example annotation document using the Graph Annotation Format (GrAF).

26 LAF takes input from several other specifications: the header files resemble the ones used in CES, which in turn are based on TEI headers. ISO 24610-1:2006 can be used for these feature structures. However, the standard recommends its own representation format shown in figure 2 as a more concise notation.

27 What is somewhat disturbing is the fact that a document grammar for the Graph Annotation Format was removed when the draft standard moved from from DIS to FDIS. The DIS version contained an XML schema file in the informative annex of the specification while the FDIS contains only fragments of a RELAX NG document grammar. Since the FDIS was approved as International Standard in 2012 without any comments regarding this topic, we assume that this is also the case for the final version.

2.4. The Syntactic Annotation Framework (SynAF)

28 The Syntactic Annotation Framework (SynAF, ISO 24615:2010) pursues the goal of defining both a meta-model for syntactic annotation and a set of data categories. In contrast to the more specific Morpho-Syntactic Annotation Framework (MAF), which is discussed in the next subsection, SynAF had already been published as an International Standard in 2010. The latest version that is publicly available for free is ISO/FDIS 24615, but an early version is discussed by Declerck (2006). SynAF is based on the Penn Treebank initiative, the Negra/Tiger initiative, and the ISST initiative and has been developed mainly by the LIRICS Consortium. While MAF deals with part of speech, morphological and grammatical features, SynAF deals with the annotation of syntactic constituency of groups of MAF word forms in sentence boundaries.

29 The meta-model for SynAF contains the generic class of Syntactic Nodes and Syntactic Edges, which together form a Syntactic Graph. Syntactic Nodes can be differentiated into T_Nodes (terminal nodes)—that is, the morpho-syntactic annotated word forms of MAF, defined over one or more spans—and NT_Nodes (non-terminal nodes of a syntax tree). The T_Nodes are annotated with syntactic data categories according to the word level, whereas the NT_Nodes are annotated with syntactic categories according to the phrase, clause, or sentence level.

30 Syntactic Edges are used to represent relations between Syntactic Nodes, such as dependency relations. The edges can be specified as primarySyntacticEdge (expressing the constituency relationship) or secondarySyntacticEdge, which "may be used to express the relationship between a head and a coreferent of its omitted dependent" (ISO/FDIS 24615, 14). Since the standard does not propose a specific tag set but only generic classes and specific data categories, there are several possible serialization formats. Romary et al. (2011) propose the <tiger2> XML format; another natural selection would be the Graph Annotation Format defined in LAF.

2.5. The Morpho-Syntactic Annotation Framework (MAF)

31 The Morpho-Syntactic Annotation Framework is closely connected to the Syntactic Annotation Framework (SynAF) discussed in the previous section. MAF is not yet an International Standard but is in the stage of an FDIS (ISO/FDIS 24611). The last version freely available to the public is ISO/CD 24611. However, the basic concepts of the specification such as the two-level structuring for tokens and word forms, and the ambiguity handling are discussed by Clément and de la Clergerie (2005).

32 MAF uses stand-off annotation as well and represents an annotated document as the primary data (called a "raw document" by Clément and de la Clergerie 2005) and a set of annotations. An input document can be divided into tokens, which can be used as anchors for word forms. Tokens resemble the regions in LAF—that is, they represent segments of the primary data. MAF does not provide an addressing schema used to refer to positions but instead relies on externally defined addressing schemas.[10]

33 Similar to LAF, these tokens can be organized in a directed acyclic graph (DAG) called a token lattice. Word forms carry the annotation by using feature structure representations and refer to tokens in an m:n-relation (where one or more tokens anchors one or more word forms). Word forms, too, can be organized—in a word form lattice. Figure 3 shows an example annotation of the sentence "I wanna put up new wallpaper."[11]

```
<maf xmlns="http://www.iso.org/ns/MAF" document="sample.txt" addressing="char_offset">
 <olac:olac
 xmlns:olac="http://www.language-archives.org/OLAC/1.0/"
 xmlns="http://purl.org/dc/elements/1.1/">
  <creator>Maik Stührenberg</creator>
 </olac:olac>
 <token xml: form="I" from="0" to="1"/>
 <token xml: join="right" form="wan" from="2" to="5"/>
 <token xml: join="left" form="na" from="5" to="7"/>
 <token xml: form="put" from="8" to="11"/>
 <token xml: form="up" from="12" to="14"/>
 <token xml: form="new" from="15" to="18"/>
 <token xml: form="wall" from="19" to="23"/>
 <token xml: form="paper" from="23" to="28"/>
 <token xml: form="." from="28" to="29">.</token>
 <wordForm lemma="I" tokens="#t1">
  <fs>
   <f name="pos">
    <symbol value="PP"/>
   </f>
  </fs>
 </wordForm>
 <wordForm lemma="want" tokens="#t2">
  <fs>
   <f name="pos">
    <symbol value="VBP"/>
   </f>
  </fs>
 </wordForm>
 <wordForm lemma="to" tokens="#t3">
  <fs>
   <f name="pos">
    <symbol value="TO"/>
   </f>
  </fs>
 </wordForm>
 <wordForm tokens="#t2 #t3"/>
 <wordForm lemma="put" tokens="#t4"/>
 <wordForm lemma="up" tokens="#t5"/>
 <wordForm lemma="put_up" tokens="#t4 #t5">
  <fs>
   <f name="pos">
    <symbol value="VB"/>
   </f>
  </fs>
 </wordForm>
 <wordForm lemma="new" tokens="#t6">
  <fs>
   <f name="pos">
    <symbol value="JJ"/>
   </f>
  </fs>
 </wordForm>
 <wordForm lemma="wallpaper" tokens="#t7 #t8">
  <fs>
   <f name="pos">
    <symbol value="NN"/>
   </f>
  </fs>
 </wordForm>
</maf>
```

Figure 3: Example annotation using MAF's current serialization format.

34 Instead of stand-off annotation, it is possible to use inline annotation for the token content; in fact, most examples in ISO/CD 24611 use this notation. In this case the value of the @from attribute would be used as element content of the <token> element and the @from and @to attributes would be omitted. However, following the standard, this is not recommended since it may conflict with other annotations.

35 The morpho-syntactic content is represented by feature structures: ISO/CD 24611 directly refers to ISO 24610-1:2006. Metadata may be included according to the OLAC metadata specification (Simons and Bird 2008) using the OLAC namespace as seen in figure 3.

36 In addition, ISO/FDIS 24611 contains a RELAX NG-like specification, some annotated examples and a list of morpho-syntactic data categories as part of its appendixes.

3. The Relation of the TEI to the Current *de jure* Standards

37 In this section the relation between the TEI and the previously mentioned standards will be discussed, focusing on aspects of their notation format and annotation models. Bański and Przepiórkowski have already stated the fact that the TEI is a direct ancestor of these standards:

> The current standards that have been or are being established by ISO TC 37 SC 4 committee …, known together as the LAF (Linguistic Annotation Framework) family of standards, … descend in part from an early application of the TEI, back when the TEI was still an SGML-based standard. That application was the Corpus Encoding Standard …, later redone in XML and known as XCES …. XCES was a conceptual predecessor of the current ISO LAF pivot format for syntactic interoperability of annotation formats, GrAF (Graph Annotation Framework) …. GrAF defines an XML serialization of the LAF data model consisting of directed acyclic graphs with annotations (also expressible as graphs), attached to nodes. This basic data model is in fact common to the TEI formats defined for the NCP, the LAF family of standards, and the other standards and best practices …. (2010b, 36)

3.1. Influence on the Data Model

38 In the field of Digital Humanities there has been the assumption that text is hierarchically structured (see, for example, Coombs et al. 1987 or the OHCO thesis postulated by DeRose et al. 1990 and Renear et al. 1996, stating that a text is an Ordered Hierarchy of Content Objects), and therefore markup languages which were developed to annotate mainly textual content use the formal model of a tree.

39 But in fact, there are several authors that tend to agree that the formal model of XML instances is that of a graph: Abiteboul et al. 2000, Polyzotis and Garofalakis 2002, Gou and Chirkova 2007, Møller and Schwartzbach 2011, and Jettka and Stührenberg 2011. In particular, the use of the XML-inherent integrity constraints—that is, ID/IDREF/IDREFS token-type attributes (in XML DTD syntax) or xs:ID/xs:IDREF/xs:IDREFS and xs:key/xs:keyref (in XSD syntax), respectively, which are supported by document grammar formalisms—can be used to represent graph structures in XML. An example for such an XML serialization of a graph can be observed in the way in which an edge in GrAF is constructed by referring to the IDs of already established nodes via the @from and @to attributes. Similar examples can be found in the XStandoff format (Stührenberg and Jettka 2009; Witt et al. 2011; Jettka and Stührenberg 2011).

40 Apart from a representation format for graphs, networks, and trees found in TEI since P3, the refined and enhanced feature structure representation format of TEI P5 has been a great step in establishing a more expressive formal model. In addition, other specifications developed for various projects, such as XStandoff, NITE (Carletta et al. 2005), or the Potsdamer Austauschformat für linguistische Annotation[12] (PAULA, Dipper et al. 2007), propagate graph-based formal models.

41 Therefore, the TEI cannot be seen as the direct or single ancestor of the current standards in development. However, it seems that this newer graph-based formal model (that is dependent on the existence of a document grammar using the aforementioned integrity constraints) may play a greater part in future XML formats (especially those for structuring multiply annotated data), and one may argue that the TEI has accompanied this change from a strictly hierarchical to a graph-based formal model.

3.2. Influence on Notation Format

42 The notation format that is used by all standards discussed here is stand-off annotation. Although stand-off annotation is not a generic TEI concept, the TEI Guidelines have long included mechanisms to deal with overlapping markup, namely milestone elements, fragmentation and reconstruction, and multiple encodings of the same information.[13] Moreover, it was the previously mentioned Corpus Encoding Standard (CES), a modification of TEI P3 that made stand-off annotation the default model for linguistic corpora. In the current version of the TEI (P5) the term "stand-off markup" is discussed in Chapters 16.9 and 20.4, firmly establishing the concept of separating primary data and markup in the wider text encoding community. This support for stand-off annotation is rated as a crucial point by Bański and Przepiórkowski: "Any standards adopted for these levels should allow for stand-off annotation, as is now common practice and as is virtually indispensable in the case of many levels of annotation, possibly involving conflicting hierarchies" (2010a, 98).

43 Although stand-off annotation can still be cumbersome to manage (especially when positions in the primary data are used to establish anchors and regions), some software products have been developed during the past years to support this notation—for example, the web-based annotation platform "Serengeti" (which uses XStandoff—see Stührenberg et al. 2007; Poesio et al. 2011) or the "Glozz Annotation Platform" (Widlöecher and Mathet 2009). Among the various candidates for dealing with multiple (and possibly overlapping) annotations, stand-off markup seems to be the most promising. (See Bański 2010 for a discussion of advantages and disadvantages of using TEI stand-off annotation.)

3.3. Influence on the Annotation Model

44 One of the building blocks of the TEI's success among various scholars is the fact that it does not define a normative standard but rather guidelines. These recommendations try to not constrain the user to a single way of encoding but leave a large amount of personal freedom (and responsibility) to the user, while other annotation formats try to be as strict as possible to reflect a certain annotation model and theory.

45 The generic markup that is manifested in the TEI's feature structure representation is informed by this permissive attitude. As a consequence, all current International Standards for linguistic data use generic elements and attributes (and especially feature structures) to store annotation information. The use of such generic markup has both advantages and disadvantages. On the one hand it helps to separate the meaning (the concept) of an annotation from its serialization (a separation introduced by Bayerl et al. 2003 and Witt 2004), establishing a basis for multiply annotated corpora. But on the other hand, a generic annotation format is generally more verbose and makes only little use of the hierarchical relations between elements inherent in XML. In addition, it relies heavily on a given set of standardized data categories to assure the comparability of annotation.

4. Conclusion

46 A comparison of the TEI Guidelines with the International Standards discussed in the previous sections leaves us with mixed results. On the one hand, the ISO specifications have the advantage of being de jure standards (at least if the standardization process will be finished for MAF). On the other hand, this status is a mixed blessing. Since International Standards are the outcome of a procedure relying on consensus, the results are often compromise-ridden. Moreover, specifications can get mired in long approval processes: LAF is a case in point, since it took so many years to reach the status of an International Standard. This long gestation raised problems for other standards, such as MAF, that refer to LAF's components even before the standard was finalized. In addition, users not familiar with the relationships between the different standards may find it difficult to keep track of specification status and dependencies. To help such users, we have developed a web-based information system presenting an overview of these relations (Stührenberg et al. 2012).

47 In contrast, the TEI Guidelines represent a stable and mature representation format for annotation. Although it is also based on consensus, by maintaining a greater variety of possible

annotation solutions it is less prone to compromise.[14] Another advantage over the standards discussed in this article is that the TEI can be used as is without the need to add further specifications, such as an external metadata format. In addition, the TEI tag set is highly modular and can be modified easily by using the web-based "Roma" tool, resulting in a strict or rich feature set depending on one's own needs. The comprehensive Guidelines themselves and a large helpful community complement these benefits. Therefore, it should not be surprising that the TEI remains a recommended annotation format for encoding linguistic corpora, following Przepiórkowski and Bański: "We conjecture that—given the stability, specificity and extensibility of TEI P5 and the relative instability and generality of some of the other proposed standards—this approach is currently the optimal way of following corpus encoding standards." (2009, 250).

48 However, with International Standards such as the Linguistic Annotation Framework, the Morpho-Syntactic Annotation Framework, and the Syntactic Annotation Framework, normative efforts to ease the exchange of linguistically annotated data are finally emerging. It will be interesting to observe the final version of MAF and especially the application of LAF and MAF in the wild.

49 Regarding the relationship between the TEI Guidelines and the discussed de jure standards, one can observe that the former may have influenced current specifications in many ways. However, especially for the data model and notation format, other projects and specifications played important roles as well.

5. Recommendations

50 Current linguistic researchers are spoiled for choice: in addition to well-established de facto standards such as the TEI, international de jure standards are on the rise. Projects such as CLARIN or FLARENET promise to help users choose among them by providing recommendations and guidelines as the aforementioned web-based information system. Apart from that, it seems that the combination of generic annotation formats such as the feature structure representation format present in the TEI P5, ISO 24610-1:2006, and ISO 24610-2:2011 and respective data category sets will be a valid candidate for a sustainable annotation format. Data categories should be registered via the official implementation of ISO 12620:2009, ISOcat, available at http://www.isocat.org.

51 A practical additional interim solution could be the setup of an ISOcat TEI data category set providing all of the elements and attributes in P5. In conjunction with a stylesheet transforming inline TEI to a stand-off TEI feature structure representation (with the respective ISOcat references), the resulting output format should be compatible with ISO 24610-1:2006 and could be used as a starting point for LAF-based annotations.

52 As a side-effect, users familiar with the TEI could use their existing annotation tool chain. Future versions of the TEI Guidelines should further embrace the noticeable trend of using stand-off notation, possibly introducing it to a broader range of linguistic researchers and even for other non-linguistic uses of the TEI.

Bibliography

Abiteboul, Serge, Peter Buneman, and Dan Suciu. 2000. *Data on the Web: From Relations to Semistructured Data and XML*. San Francisco: Morgan Kaufman.

Piotr Bański. 2010. "Why TEI standoff annotation doesn't quite work: and why you might want to use it nevertheless." In *Proceedings of Balisage: The Markup Conference*, 2010. Vol. 5 of Balisage Series on Markup Technologies. doi:10.4242/BalisageVol5.Banski01.

Bański, Piotr, and Adam Przepiórkowski. 2010a. "TEI P5 as a Text Encoding Standard for Multilevel Corpus Annotation." In *Digital Humanities 2010 Conference Abstracts*, 98–100. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-616.pdf.

———. 2010b. "The TEI and the NCP: the Model and its Application." In *Proceedings of LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, 34–38. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf.

Bayerl, Petra Saskia, Harald Lüngen, Daniela Goecke, Andreas Witt, and Daniel Naber. 2003. "Methods for the Semantic Analysis of Document Markup." In *Proceedings of the 2003 ACM Symposium on Document Engineering*, 161–170. New York: ACM.

Broeder, Daan, Oliver Schonefeld, Thorsten Trippel, Dieter van Uytvanck, and Andreas Witt. 2011. "A Pragmatic Approach to XML Interoperability – the Component Metadata Infrastructure (CMDI)." In *Proceedings of Balisage: The Markup Conference 2011*. Vol. 7 of Balisage Series on Markup Technologies. doi:10.4242/BalisageVol7.Broeder01.

Carletta, Jean, Stefan Evert, Ulrich Heid, and Jonathan Kilgour. 2005. "The NITE XML toolkit: data model and query language." *Language Resources and Evaluation* 39 (4): 313–334.

Clément, Lionel, and Èric Villemonte de la Clergerie. 2005. "MAF: A Morphosyntactic Annotation Framework." In *Proceedings of the 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 90–94. Poznań, Poland: Wydawnictwo Poznańskie.

Coombs, James H. Allen H. Renear, and Steven J. DeRose. 1987. "Markup Systems and the Future of Scholarly Text Processing." *Communications of the ACM* 30 (11): 933–947.

Dalby, David, Lee Gillam, Christopher Cox, and Debbie Garside. 2004. "Standards for Language Codes: Developing ISO 639." In *LREC 2004: Fourth International Conference on Language Resources and Evaluation*, 127–130. Paris: ELRA.

Declerck, Thierry. 2006. "SynAF: Towards a Standard for Syntactic Annotation." In *Book of Abstracts* [conference abstracts from LREC 2006], 229–232. Paris: ELRA.

Declerck, Thierry, Nancy Ide, and Thorsten Trippel. 2007. "Interoperable Language Resources." *SDV – Sprache und Datenverarbeitung* 31 (01/02): 101–113.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What is text, really?" *Journal of Computing in Higher Education* 1 (2): 3–26.

Dipper, Stefanie, Michael Götze, Uwe Küssner, and Manfred Stede. 2007. "Representing and Querying Standoff XML." In *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Data Structures for Linguistic Resources and Applications*, edited by Georg Rehm, Andreas Witt, and Lothar Lemnitzer, 337–346. Tübingen: Gunter Narr.

Gou, Gang, and Rada Chirkova. 2007. "Efficiently Querying Large XML Data Repositories: A Survey." *IEEE Transactions on Knowledge and Data Engineering* 19 (10): 1381–1403.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. "XCES: An XML-based Encoding Standard for Linguistic Corpora." In *Second International Conference on Language Resources and Evaluation*, 825–830. Paris: European Language Resources Association.

Ide, Nancy. 1998. "Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora". In *First International Conference on Language Resource and Evaluation*, 463–470. Paris: ELRA.

Ide, Nancy, and Laurent Romary. 2007. "Towards International Standards for Language Resources." In *Evaluation of Text and Speech Systems*, edited by Laila Dybkjaer, Holmer Hemsen, and Wolfgang Minker, 263–284. Dordrecht: Springer.

International Organization for Standardization/International Electrotechnical Commission. 2012. "ISO/IEC Directives, Part 1: Procedures for the technical work." 9th Edition, March 8, 2012. http://isotc.iso.org/livelink/livelink?func=ll&objId=10563026&objAction=Open&nexturl=%2Flivelink%2Flivelink%3Ffunc%3Dll%26objId%3D4230455%26objAction%3Dbrowse%26sort%3Dsubtype.

Jettka, Daniel, and Maik Stührenberg. 2011. "Visualization of concurrent markup: From trees to graphs, from 2D to 3D." In *Proceedings of Balisage: The Markup Conference 2011*. Vol. 7 of Balisage Series on Markup Technologies. doi:10.4242/BalisageVol7.Jettka01.

Langendoen, D. Terence, and Gary F. Simons. 1995. "A Rationale for the TEI Recommendations for Feature-Structure Markup." *Computers and the Humanities* 29 (3): 191–209.

Monachini, Monica, Valeria Quochi, Nicoletta Calzolari, Núria Bel, Gerhard Budin, Tommaso Caselli, Khalid Choukri, Gil Francopoulo, Erhard Hinrichs, Steven Krauwer, Lothar Lemnitzer, Joseph Mariani, Jan Odijk, Stelios Piperidis, Adam Przepiorkowski, Laurent Romary, Helmut Schmidt, Hans Uszkoreit, and Peter Wittenburg. 2011. "The Standards' Landscape Towards an Interoperability Framework: The FLaReNet proposal Building on the CLARIN Standardisation Action Plan." http://www.flarenet.eu/sites/default/files/FLaReNet_Standards_Landscape.pdf.

Møller, Anders, and Michel I. Schwartzbach. 2011. "XML Graphs in Program Analysis." *Science of Computer Programming* 76 (6): 492–515.

Poesio, Massimo, Nils Diewald, Maik Stührenberg, Jon Chamberlain, Daniel Jettka, Daniela Goecke, and Udo Kruschwitz. 2011. "Markup Infrastructure for the Anaphoric Bank: Supporting Web Collaboration." In *Modeling, Learning and Processing of Text Technological Data Structures*, edited by Alexander Mehler, Kai-Uwe Kühnberger, Henning Lobin, Harald Lüngen, Angelika Storrer, and Andreas Witt, 197–218. Berlin: Springer.

Pollard, Carl, and Ivan A. Sag. 1987. *Information-based Syntax and Semantics*. Menlo Park: CSLI.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: The University of Chicago Press.

Polyzotis, Neoklis, and Minos Garofalakis. 2002. "Statistical Synopses for Graph-Structured XML Databases." In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, 358–369. New York: ACM.

Przepiórkowski, Adam, and Piotr Bański. 2009. "Which XML Standards for Multilevel Corpus Annotation?" http://bach.ipipan.waw.pl/~adamp/Papers/2009-ltc-tei/ltc-030-przepiorkowski.pdf.

Renear, Allen H., Mylonas, Elli, and David D. Durand. 1996. "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies." *Selected Papers from the ALLC/ACH Conference, Christ Church, Oxford, April 1992*. Vol. 4 of Research in Humanities Computing. 263–280.

Romary, Laurent, Amir Zeldes, and Florian Zipser. 2011. "<tiger2/> – Serialising the ISO SynAF Syntactic Object Model." *Computing Research Repository (CoRR)*. http://arxiv.org/pdf/1108.0631v1.

Simons, Gary F. 2007. "Linguistics as a community activity: The paradox of freedom through standards." In *Time and Again: Theoretical and Experimental Perspectives on Formal Linguistics: Papers in Honor of D. Terence Langendoen*, edited by William D. Lewis, Simin Karimi, Heidi Harley, and Scott Farrar, 235–250. Amsterdam: John Benjamins.

Simons, Gary F., and Steven Bird. 2008. "OLAC Metadata." *Open Language Archives Community Standard*. http://www.language-archives.org/OLAC/metadata-20080531.html.

Stegmann, Jens, and Andreas Witt. 2009. "TEI Feature Structures as a Representation Format for Multiple Annotation and Generic XML Documents." *Proceedings of Balisage: The Markup Conference*. Balisage Vol. 3 of Series on Markup Technologies. doi:10.4242/BalisageVol3.Stegmann01.

Stührenberg, Maik, Daniela Goecke, Nils Diewald, Irene Cramer, and Alexander Mehler. 2007. "Web-based Annotation of Anaphoric Relations and Lexical Chains." In *Proceedings of the Linguistic Annotation Workshop*, 140–147. http://www.aclweb.org/anthology/W/W07/W07-1523.pdf.

Stührenberg, Maik, and Daniel Jettka. 2009. "A Toolkit for Multi-dimensional Markup: The Development of SGF to XStandoff." *Proceedings of Balisage: The Markup Conference* 2009. Vol. 3 of Balisage Series on Markup Technologies. doi:10.4242/BalisageVol3.Stuhrenberg01.

Stührenberg, Maik, Antonina Werthmann, and Andreas Witt. 2012. "Guidance through the Standards Jungle for Linguistic Resources." In *Proceedings of the LREC 2012 Workshop on Collaborative Resource Development and Delivery*, 9–13.

Thompson, Henry S., and David McKelvie. 1997. "Hyperlink Semantics for Standoff Markup of Read-only Documents." In *Proceedings of SGML Europe '97: The next decade – Pushing the Envelope*, 227–229.

Widlöcher, Antoine, and Yann Mathet. 2009. "La plate-forme Glozz : environnement d'annotation et d'exploration de corpus". In *Actes de la 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009) – Session posters*. http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_120.pdf.

Witt, Andreas. 2004. "Multiple Hierarchies: New Aspects of an Old Solution." *Proceedings of Extreme Markup Languages, Montréal*.
http://conferences.idealliance.org/extreme/html/2004/Witt01/EML2004Witt01.html.

Witt, Andreas, Daniela Goecke, Maik Stührenberg, and Dieter Metzing, 2011. "Integrated Linguistic Annotation Models and Their Application in the Domain of Antecedent Detection". In *Modeling, Learning and Processing of Text Technological Data Structures*, edited by Alexander Mehler, Kai-Uwe Kühnberger, Henning Lobin, Harald Lüngen, Angelika Storrer, and Andreas Witt, 197–218. Berlin: Springer.

Notes

1 See the projects' websites at http://www.clarin.eu/ and http://www.flarenet.eu/, respectively, for further information.

2 The website located at http://www.tc37sc4.org/ provides some further information.

3 P-members are contrasted with O-members, who only observe but still have the right to comment on the process.
4 If no negative votes are cast the DIS proceeds to the publication stage immediately.
5 See Dalby et al. (2004) for further details about the design philosophy of this special standard.
6 Apart from the specifications discussed in this section there are of course other standards that may be of interest, such as the Lexical Markup Framework (LMF, ISO 24613:2008). However, due to space restrictions we limit the discussion to the annotation formats described in this article. We will not discuss in detail any metadata standards, such as ISO 12620:2009 (Data Category Registry, DCR), which can be used together with generic annotation formats to provide further semantics for a linguistically encoded text.
7 For an overview of HPSG, see Pollard and Sag (1987, 1994).
8 See Langendoen et. al (1995) for a discussion of the TEI recommendations for feature structure markup.
9 See http://www.isocat.org for more information about both ISO 12620:2009 and about the ISOcat registry.
10 The current version of MAF includes the notion, that “character offsets may be sufficient” in the simplest case.
11 The original example was taken from http://korpling.german.hu-berlin.de/tiger2/homepage/tiger1.html and was adapted to meet further MAF requirements.
12 Potsdam Interchange Format for Linguistic Annotation.
13 Early usage of stand-off annotation can be found in the second phase of the TIPSTER project in 1996. A discussion of the concept can be found in Thompson and McKelvie (1997). The P3 version of the TEI did not include the term stand-off as such but supported the connection of analytic and interpretive markup outside of textual markup and embedded markup (Chapter 14.9). The current P5 includes a whole chapter dealing with stand-off markup (Chapter 16.9).
14 One has to admit that one of the disadvantages of the TEI is the fact that it frequently allows too many ways of annotating a certain text feature. This can also be seen as a limiting compromise.

Cite this article

Electronic reference

Maik Stührenberg, « The TEI and Current Standards for Structuring Linguistic Data », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 15 October 2012, connection on 05 November 2012. URL : http://jtei.revues.org/523 ; DOI : 10.4000/jtei.523

Author

Maik Stührenberg
Maik Stührenberg received his Ph. D. in Computational Linguistics and Text Technology from Bielefeld University in 2012. After graduating in 2001, he worked on various projects at Justus-Liebig-Universität Gießen, Bielefeld, and at the Institut für Deutsche Sprache (IDS, Institute for the German Language) in Mannheim. He is currently employed as a research assistant at Bielefeld University and is involved in NA 105-00-06 AA, the German mirror committee of ISO TC37 SC4. His main research interests include specifications for structuring multiply annotated data (especially linguistic corpora), query languages, and query processing.

Abstract

The TEI has served for many years as a mature annotation format for corpora of different types, including linguistically annotated data. Although it is based on the consensus of a large community, it does not have the legal status of a standard. During the last decade, efforts have been undertaken to develop definitive *de jure* standards for linguistic data that not only act as a normative basis for the exchange of language corpora but also address recent advancements in technology, such as web-based standards, and the use of large and multiply annotated corpora. In this article we will provide an overview of the process of international standardization and discuss some of the international standards currently being developed under the auspices of ISO/TC 37, a technical committee called "Terminology and other Language and Content Resources". After that the relationship between the TEI Guidelines and these specifications, according to their formal model, notation format, and annotation model, will be discussed. The conclusion of the paper provides recommendations for dealing with language corpora.

Index terms

Keywords : Standards, ISO/TC 37/SC 4, Feature Structures, Linguistic Annotation Framework (LAF), Morpho-Syntactic Annotation Framework (MAF), Syntactic Annotation Framework (SynAF)