Published in: Ide, Nancy/Baker, Collin/Fellbaum, Christiane/Passonneau, Rebecca (eds.): Proceedings of the LREC-12 Workshop on Collaborative Resource Development and Delivery. Istanbul, Turkey, May 2012. - European Language Resources Association (ELRA), 2012. pp. 9-13.

Guidance through the standards jungle for linguistic resources

Maik Stührenberg, Antonina Werthmann, Andreas Witt

Institut für Deutsche Sprache
R 5, 6-13
D-68161 Mannheim
{stuehrenberg|werthmann|witt}@ids-mannheim.de

Abstract

Research today is often performed in collaborated projects composed of project partners with different backgrounds and from different institutions and countries. Standards can be a crucial tool to help harmonizing these differences and to create sustainable resources. However, choosing a standard depends on having enough information to evaluate and compare different annotation and metadata formats. In this paper we present ongoing work on an interactive, collaborative website that collects information on standards in the field of linguistics as a means to guide interested researchers.

1. The importance of standards for collaborated resource development

Research today is often performed by teams of project partners from different institutions and countries. The first steps in such projects often focus on architectural issues, such as the choice of annotation formats or metadata standards. Project partners can only choose the best standards for their projects, however, when they have enough information to evaluate and compare standards. In this paper we will present ongoing work on an interactive, collaborative website that collects information on standards in the field of linguistics.

2. Different views on standards

Over the last 20 years, the annotation of linguistic phenomena has gone through a number of transitions, on both a general "meta" level and a more specific application-oriented level. First, meta languages such as SGML and later XML were established as standards. These two meta languages replaced the proprietary and binary formats that were used in annotation projects for linguistic data and were developed by the ISO/IEC (in case of SGML) and the W3C (in case of XML). Both organizations act in a wide field of specifications that may affect linguistic research, such as the W3C Recommendations XPath, XSLT, XML Schema or the Internationalization Tag Set (Lieske and Sasaki, 2007), or the ISO standards RELAX NG or Schematron. In addition, other general standards that are also crucial for language resources were developed by other organizations such as Unicode (The Unicode Consortium, 1991). These various specifications laid the groundwork for the application-oriented level, where initial steps were undertaken to harmonize the various efforts of linguistic researchers by developing a unified tagset for linguistic annotation. This was necessary since use of the same underlying meta language did not guarantee easy exchange of data or a sustainable use of the meta language (Stührenberg, 2008). One result of this movement was the Text Encoding Initiative (TEI) and its Guidelines. Development of TEI began in 1987 as an SGML application and the latest XML-based version, P5, was released in 2007 (Burnard and Bauman, 2007) and updated 2011 (Burnard and Bauman, 2011). It comprises 22 modules of over 520 elements and over 430 attributes, and allow for the annotation of various linguistic phenomena. Since the TEI is quite complex but has certain shortcomings regarding some linguistic theories, a third major transition regarding annotation of linguistic corpora is taking place.

There are already numerous specifications that deal with various aspects of linguistic annotation. Amongst these are the SGML-based Corpus Encoding Standard CES (Ide and Priest-Dorman, 1996; Ide, 1998), which has been developed within the Expert Advisory Group on Language Engineering Standards (EAGLES) as an application of the TEI P3 (Expert Advisory Group on Language Engineering Standards, 1996), and its XML-based successor XCES (Ide et al., 2000). Following the work of the EAGLES initiative, the ISLE (International Standards for Language Engineering) project, which has been carried out in collaboration between American and European groups under the Human Language Technology (HLT) programme within the EU-US International Research Co-operation, continued to develop and promote language technology standards, guidelines and tools (Calzolari et al., 2002).

Other annotation formats and frameworks have been developed through the course of several research projects, including the Potsdam exchange format for linguistic annotation (Potsdamer Austauschformat für Linguistische Annotationen, PAULA) (Dipper, 2005) or the Sekimo Generic Format (SGF) (Stührenberg and Goecke, 2008) and its successor XStandoff (Stührenberg and Jettka, 2009).

Since 2005, at least half a dozen efforts to standardize (technically, to create ISO standards for) various aspects of linguistic researches have been attempted. Among these specifications are the general Feature Structures (ISO/TC 37/SC 4, 2006) and the Linguistic Annotation Framework (ISO/TC 37/SC 4, 2011), the more specific Morpho-Syntactic Annotation Framework (ISO/TC 37/SC 4, 2008), the Syntactic Annotation Framework (ISO/TC 37/SC 4, 2010), and the Data Category Registry (DCR) (ISO/TC 37/SC 3, 2004), to name just the most prominent. The most recent (i.e. final) versions of these standards are usually not open and freely available on the Internet (although libraries often grant access to the public). Some information can be derived from scientific articles but these may already be out of date. Although most

of the standards mentioned above do relate to each other, the standardization process has no mechanism to coordinate standards, which may result in specifications becoming out of sync. Another practical issue is choosing which conceptual layer is covered by the standard (e.g. syntax, semantic, etc.).

2.1. Technical aspects

Technical questions such as the grammar formalism used or the notation can have direct consequences for choosing tools to process annotated resources. Some specifications deal with a single layer (such as the Morpho-Syntactic Annotation Framework and the Syntactic Annotation Framework), while others provide a general framework such as the Linguistic Annotation Framework. Others are not used for direct annotation at all; one example is the Data Category Registry, which should only be used as a registry for annotation standards concepts.

Most of the current annotation standards use the concept of standoff annotation introduced (Thompson and McKelvie, 1997) and discussed in the TEI as well. As a result, it is necessary to find/create annotation tools capable of dealing with the separation of content and markup, limiting the choice of tools that can be used to annotate resources – although one may observe that support for standoff annotation has increased in recent years (e.g. the web-based Serengeti annotation tool (Stührenberg et al., 2007), the Glozz Annotation Platform (Widlöcher and Mathet, 2009; Mathet and Widlöcher, 2011) or the newly developed Slate (Kaplan et al., 2011)).

2.2. Formal aspects

Among the formal aspects are the formal model, the constraint language used to define the markup language (and its respective expressive power), and the annotation model (inline vs. standoff). Although the formal model of an XML instance is that of a single-rooted tree, it is possible to encode graphs in XML as well (one has to differentiate between the XML instance as such which forms a tree and the language that is represented by it, which has no further restrictions). This can be achieved by using either quite general frameworks, such as the Linguistic Annotation Framework or Feature Structures, or by using meta markup languages, such as XStandoff.

The aspect of the constraint language used may be of interest regarding the expressive power of the markup language. This expressivity can be compared both in terms of technical features (such as data typing) and formal power. Both aspects have been subject to different research, e.g. (Murata et al., 2005) built up a taxonomy of schema languages which was refined by (Stührenberg and Wurm, 2010).

3. Providing Guidance

A large number of standards can be used in the creation of sustainable linguistic resources. Within the CLARIN-D project, the IDS is responsible for providing insight into various aspects of linguistic standards. The work presented in this paper aims to help interested researchers understand the relationship between various specifications and to choose the right standard for a given task. To support this, we are developing a lightweight and transparent taxonomy that can be used as an online guide for the most recent (and most prominent) specifications for language resources, especially annotation of linguistic data. This online guide will feature information addressing the issues raised here to help researchers differentiate between standards and choose the right one. It consists of two parts. The first part contains lightweight XML metadata descriptions of the various standards. This data is in the form of stripped-down markup language that can be easily modified with a text editor. The metadata is coded based on the TEI header, while the description of the features (including the aforementioned technical and formal aspects) is coded based on TEI's feature structures which in turn was standardized as (ISO/TC 37/SC 4, 2006). Following the distinction between technical and formal aspects we make assumptions about the meta language used (SGML vs XML), the constraint language that defines the markup language and the respective grammar class, and the notation (inline vs standoff), amongst others.

The second part takes these lightweight XML metadata descriptions as a knowledge base and allows the filtering of this data according to the different criteria stated above. The results can be transformed into different output formats that are readable by web browsers, and include textual and graphical representations.

The main parts of this system are the XML descriptions of annotation formats (or other standards), a database that stores the annotation instance (e.g. a native XML database) and a web frontend for both input and output using stylesheet transformation. The web frontend is designed to make the system useful for projects with many partners. We have currently completed both the XML annotation format and prototypic instances of the specifications' description (an excerpt is shown in Listing 1).

Listing 1: Example of a specification description

```
<spec xmlns:xsi="http://www.w3.org/2001/XMLSchema-
     instance" xml:id="SpecXces" topicRef="TopicGenAnn"
 xsi:noNamespaceSchemaLocation="http://localhost:8080/
      exist/apps/clarin/xsd/spec.xsd">
 <titleStmt>
  <title>XCES: Corpus Encoding Standard in XML</title>
 </titleStmt>
 <scope>Corpus annotation</scope>
 <description>
  <p>XCES is the XML version of the CES (Corpus Encoding Standard) ... </p>
  <!--[...]-->
 </description>
 <version xml:id="SpecXces104">
  <versionNumber>1.0.4</versionNumber>
  <date>2008-06-20</date>
  <respStmt>
   <resp>Editor</resp>
   <name type="person">Nancy Ide</name>
   <name type="person">Patrice Bonhomme</name>
  </respStmt>
  <features>
   <fs>
    <f name="metaLanguage">
     <symbol value="XML"/>
    </f>
    <f name="constraintLanguage">
     <symbol value="XSD"/>
    </f>
    <f name="grammarClass">
     <symbol value="LTG"/>
    </f>
    <f name="formalModel">
     <symbol value="Graph"/>
    </f>
    <f name="notation">
     <symbol value="Standt off"/>
```

```
    </f>
    <f name="multipleHierarchies">
     <fs>
      <f name="support">
       <binary value="yes"/>
      </f>
      <f name="item">
       <vColl>
        <string>standoff annotation</string>
       </vColl>
      </f>
     </fs>
    </f>
   </fs>
  </features>
  <address type="URL">http://www.xces.org/</address>
  <relation target="SpecCes" type="isVersionOf">
   <p>XCES is the XML instantiation of CES.</p>
  </relation>
 </version>
</spec>
```

The description of a specification can be subdivided into respective versions to distinguish different feature sets. For example, the P3 version of the TEI used the SGML meta language while form P4 onwards XML was used. However, while P4 used XML DTDs as constraint language the current P5 is based on RELAX NG. Since we only provide a small subset of any feature that can be relevant for a project, the description of the feature set is done via a TEI feature structure-like representation. Relations between specifications are described via the `relation` element. It contains two required attributes, `target` and `type`. While the former specifies the standard this one is related to, the value of the latter classifies the type of relation. We provide a list of relation types based on the DCMI Metadata Terms (DCMI Usage Board, 2010), such as *isApplicationOf* or *isVersionOf*, amongst others. DCR categories which can be obtained via ISOcat[1] could be used as well.

An even more lightweight format is used to store and describe the topics which are subsumed in a single XML instance.

Listing 2: Example of a topic description

```
<topics xmlns:xsi="http://www.w3.org/2001/XMLSchema-
     instance"
 xsi:noNamespaceSchemaLocation="http://localhost:8080/
      exist/apps/clarin/xsd/spec.xsd">
 <!--[...]-->
 <topic xml:id="TopicMetadata">
  <titleStmt>
   <title>Metadata</title>
  </titleStmt>
  <description>
   <p>Metadata contains information about other data ... </p>
   <!--[...]--A
  </description>
 </topic>
```

The format as such is defined by an XML schema description (XSD) because of XSD's strong data typing support. At present, the implementation shown above is stored into the native XML database eXist.[2] XQuery scripts transform the given information into different XHTML output files based on interactive web forms created with XForms (Boyer, 2009). Figure 1 shows a partial screenshot of the current implementation.[3]

A future incarnation will support a graphical overview of the relations between different specifications based on Scalable Vector Graphics (SVG), the one shown in Figure 1 contains a preliminary mockup. At the time of writing, the proposed system is a work in progress. The complete site will be launched, to coincide with the conference.

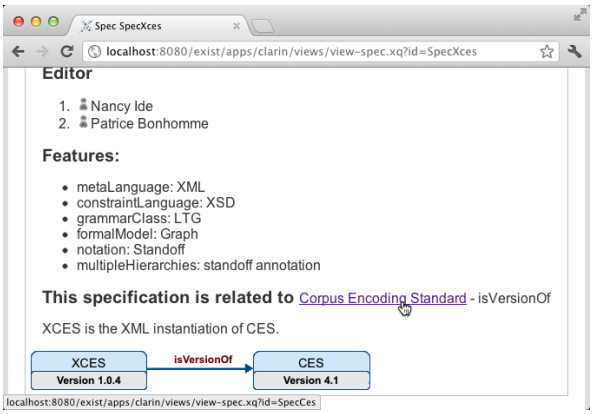

Figure 1: Partial screenshot of the current prototype.

4. Related approaches

There are already similar and related initiatives that try to help researchers deal with the variety of standards and language resources and that should be mentioned, although they cover a wider range of tools than our approach does. The LRE Map of Language Resources and Tools by FLaReNet (Fostering Language Resources Network)[4] and ELRA (European Language Resources Association) which was introduced at the LREC 2010 conference and collects information on both existing and newly-created language resources[5]. As a next step, the Language Library (Calzolari et al., 2011a) has launched for LREC 2012.

The FLaReNet Databook[6] comprises a picture of the current state of language resource technology and includes a practical orientation for the current standards landscape (Calzolari et al., 2011c; Monachini et al., 2011). Since the Databook states that information about standards has to be "constantly/periodically revised and updated by the community itself", we think that a open, web-based approach may be a means to this goal.

5. Outlook and further possible enhancements

Up until now, the relations between the specifications described are quite basic (cf. Section 3.). Possible future enhancements should not only address a more detailed graphical rendering of the relations but should enhance the type of relations as well, including mutually dependent relations between standards.

[1] See `http://www.isocat.org` for further details.

[2] See `http://www.exist-db.org` for further details.

[3] The prototype can be observed at `http://clarin.ids-mannheim.de/standards`.

[4] FLaReNet is a project initiative funded by the European Commission in the framework of the eContentplus Programme. See `http://www.flarenet.eu` for further details.

[5] A beta version can be found at `http://www.resourcebook.eu/LreMap/faces/views/resourceMap.xhtml`.

[6] Cf. `http://www.flarenet.eu/?q=FLaReNet_Databook` for further information.

6. References

John M. Boyer. 2009. XForms 1.1. W3C Recommendation, World Wide Web Consortium (W3C).

Lou Burnard and Syd Bauman, editors. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. published for the TEI Consortium by Humanities Computing Unit, University of Oxford, Oxford, Providence, Charlottesville, Bergen, 10. Version 1.0.0. Last updated on October 28st 2007.

Lou Burnard and Syd Bauman, editors. 2011. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Published for the TEI Consortium by Humanities Computing Unit, University of Oxford, Oxford, Providence, Charlottesville, Bergen, 3. Version 1.9.1. Last updated on March 5th 2011.

Nicoletta Calzolari, Alessandro Lenci, Francesca Bertagna, and Antonio Zampolli. 2002. Broadening the scope of the EAGLES/ISLE lexical standardization initiative. In *COLING-02: Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization*, pages 1–8.

Nicoletta Calzolari, Riccardo Del Gratta, Francesca Frontini, and Irene Russo. 2011a. The language library: Many layers, more knowledge. In Calzolari et al. (Calzolari et al., 2011b), pages 93–97.

Nicoletta Calzolari, Toru Ishida, Stelios Piperidis, and Virach Sornlertlamvanich, editors. 2011b. *Proceedings of the IJCNLP 2011 Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, Chiang Mai, Thailand, 11. Asian Federation of Natural Language Processing.

Nicoletta Calzolari, Monica Monachini, and Valeria Quochi. 2011c. Interoperability framework: The FLaReNet action plan proposal. In Calzolari et al. (Calzolari et al., 2011b), pages 41–49.

DCMI Usage Board. 2010. DCMI Metadata Terms. DCMI Recommendation, Dublin Core Metadata Initiative, 10.

Stefanie Dipper. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin.

Expert Advisory Group on Language Engineering Standards. 1996. EAGLES Guidelines.

Nancy M. Ide and Greg Priest-Dorman. 1996. Corpus Encoding Standard (CES). Technical report, Expert Advisory Group on Language Engineering Standards (EAGLES).

Nancy M. Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Language Resources and Evaluation (LREC 2000)*, pages 825–830, Athens, 5. European Language Resources Association (ELRA).

Nancy M. Ide. 1998. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In *Proceedings of the First International Language Resources and Evaluation (LREC 1998)*, pages 463–470, Granada, Spain. European Language Resources Association (ELRA).

ISO/TC 37/SC 3. 2004. Terminology and other language and content resources — Data categories — Part 1: Specification of data categories and management of a data category registry for language resources. Committee Draft ISO/CD 12620-1 (N 509), International Organization for Standardization, Geneva, 7.

ISO/TC 37/SC 4. 2006. Language Resource Management — Feature Structures – Part 1: Feature Structure R. International Standard ISO 24610-1:2006, International Organization for Standardization, Geneva.

ISO/TC 37/SC 4. 2008. Language Resource Management — Morpho-syntactic annotation framework. Draft International Standard ISO/DIS 24611, International Organization for Standardization, Geneva.

ISO/TC 37/SC 4. 2010. Language Resource Management — Syntactic annotation framework (SynAF). International Standard ISO 24615, International Organization for Standardization, Geneva.

ISO/TC 37/SC 4. 2011. Language Resource Management — Linguistic annotation framework (LAF). Final Draft International Standard ISO/FDIS 24612, International Organization for Standardization, Geneva, 8.

Dain Kaplan, Ryu Iida, Kikuko Nishina, and Takenobu Tokunaga. 2011. Slate — a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, 26(2):89–101.

Christian Lieske and Felix Sasaki. 2007. Internationalization Tag Set (ITS). W3C Recommendation, World Wide Web Consortium (W3C), 4.

Yann Mathet and Antoine Widlöcher. 2011. Stratégie d'exploration de corpus multi-annotés avec glozzql. In Mathieu Lafourcade and Violaine Prince, editors, *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, pages 143–148, Montpellier. Association pour le Traitement Automatique des langues (ATALA).

Monica Monachini, 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, 7.

Makoto Murata, Dongwon Lee, Murali Mani, and Kohsuke Kawaguchi. 2005. Taxonomy of XML Schema Languages Using Formal Language Theory. *ACM Transactions on Internet Technology*, 5(4):660–704.

Maik Stührenberg and Daniela Goecke. 2008. SGF – an integrated model for multiple annotations and its application in a linguistic domain. In *Proceedings of Balisage: The Markup Conference*, volume 1 of *Balisage Series on Markup Technologies*, Montréal, Québec, 8.

Maik Stührenberg and Daniel Jettka. 2009. A toolkit for multi-dimensional markup: The development of SGF to XStandoff. In *Proceedings of Balisage: The Markup Conference*, volume 3 of *Balisage Series on Markup Technologies*, Montréal, Québec, 8.

Maik Stührenberg and Christian Wurm. 2010. Refining the

Taxonomy of XML Schema Languages. A new Approach for Categorizing XML Schema Languages in Terms of Processing Complexity. In *Proceedings of Balisage: The Markup Conference*, volume 5 of *Balisage Series on Markup Technologies*, Montréal, Québec, 8.

Maik Stührenberg, Daniela Goecke, Nils Diewald, Irene Cramer, and Alexander Mehler. 2007. Web-based annotation of anaphoric relations and lexical chains. In Branimir Boguraev, Nancy M. Ide, Adam Meyers, Shigeko Nariyama, Manfred Stede, Janyce Wiebe, and Graham Wilcock, editors, *Proceedings of the Linguistic Annotation Workshop*, pages 140–147, Prague.

Maik Stührenberg. 2008. Sustainability of Text-Technological Resources. In Andreas Witt, Georg Rehm, Thomas Schmidt, Khalid Choukri, and Lou Burnard, editors, *Proceedings of the LREC 2008 Workshop "Sustainability of Language Resources and Tools for Natural Language Processing"*, pages 33–40. ELRA/ELDA.

The Unicode Consortium. 1991. The Unicode Standard. Version 1.0, Volume 1. Technical report, The Unicode Consortium, Reading, MA.

Henry S. Thompson 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*, pages 227–229, Barcelona.

Antoine Widlöcher and Yann Mathet. 2009. La plate-forme glozz : environnement d'annotation et d'exploration de corpus. In Mathieu Lafourcade and Violaine Prince, editors, *Actes de la 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009) – Session posters*, Senlis. Association pour le Traitement Automatique des langues (ATALA).