Erschienen in: ALLCACH98, Joint Conference of the ALLC and ACH, Debrecen, 1998. - Debrecen, 1998. S. 170-174.

TEI-based XML-Applications: Transcriptions

ANDREAS WITT
Linguistics

KEYWORDS: TEI XMLTranscriptions

1. TEI-BASED XML-APPLICATIONS: TRANSCRIPTIONS INTRODUCTION

1. 1. *THE STATE OF XML*

The Standard Generalized Markup Language (SGML) was defined in 1986 as an international standard (ISO 8879).[4] Its tremendous and steadily growing popularity began with its best known application HTML (in 1989) and it is intimately related to the development of the Word Wide Web (WWW; in 1991). In 1997, the World Wide Web

Consortium (W3C) proposed the eXtensible Markup Language (XML;[1]) as new language for the WWW. Unlike HTML, XML is not an *application* of SGML. It allows for the user-friendly and easy definition of SGML-applications. With XML, which is simplified SGML, it will be possible to provide documents with domain-specific markup via the WWW[1].

1. 2. THE STATE OF TEI

The significance of SGML for the annotation of technical data is quite obvious. But every subject area can benefit from standardized preparation. For applications in the humanities, the Text Encoding Initiative (TEI) defined an SGML-based encoding scheme. The efforts of the TEI manifest themselves in the Guidelines for Electronic Text Encoding and Interchange (TEI P3) which were published in 1994. [8] TEI P3 defines tag-sets for texts of various domains of the humanities, for instance prose texts, poetry, and critical editions. This paper concentrates on the (base-)tag-set for the transcription of speech, although, the approach described is applicable to all parts of the TEI encoding scheme.

1. 3. THE STATE OF TRANSCRIPTIONS

Transcriptions of speech have been introduced as a tool of research in a number of areas (e.g. psychology, anthropology). Even today, conventions for transcribing aspects of spoken language are quite heterogeneous[2]. On the one hand, differences of notation cannot be avoided because different domains using transcriptions focus on different phenomena. On the other hand, a large variety and number of phenomena must be encoded in all domains but are transcribed differently (e. g. pauses). This leads to problems when researchers want to read papers containing transcriptions. First of all, they need to familiarize themselves with the conventions used to transcribe the utterances. Furthermore, the fact that every encoding scheme that is used must be documented leads to a large amount of redundant information getting published.

2. ACCEPTANCE OF THE ENCODING SCHEME „TRANSCRIPTION OF SPEECH"

Chapter 11 of the TEI-Guidelines describes a set of tags which are very well suited for the transcription of speech. This set could be the lingua franca of all corpora of spoken language. The present situation is, unfortunately, far from this goal.
Example: The Project "Verbmobil" is „a long-term (1993 - 2000) interdisciplinary Language Technology (especially Machine Translation) project".
Verbmobil is a cooperation of four industrial and seventeen academic partners, who developed their own method of transcribing speech.[6],[3] A proposal to use the TEI-encoding scheme [9] was not realized.

One reason for not employing the TEI-encoding scheme was that it is both too hard to produce and too hard to read. TEI is not actually hard to create (if someone uses appropriate editors) but – like every encoding scheme - it is definitely hard to read. One advantage of the use of SGML in connection with style-sheets is that it allows documents to be presented in a pleasing way. Another advantage is, that the existence of SGML-parsers allows checking the correctness and unambiguousness of the encoding. The eXtensible Markup Language defines a restriction on SGML-documents which will probably lead to more widespread availability of such programs. Especially WWW-applications will be developed in the near future. Furthermore, since several special SGML features are not allowed in XML, XML is easier to learn.

3. TEI REVISITED

3. 1. PROBLEMS OF TEI P3

TEI P3 is very general. It defines a tag-set for different kinds of texts used in the humanities. With this tag-set a lot of specific phenomena can be annotated. If someone does not need a detailed level of annotations, (s)he can use only a subset of the predefined elements. If the application requires an even more sophisticated level of annotation, (s)he can extend the tag set. Because of its structure the TEI-DTD is a suitable tool for such operations. The disadvantage, however, is that it has become enormously complex. Normally, the user of the DTD does not need to worry about this complexity. Unfortunately, however, some SGML-software has difficulties to manage or simply cannot process DTD's of such complexity.

3. 2. TEI-LITE

Because of these problems, the TEI defined a more simple annotation scheme called TEI-Lite [2]. This DTD provides more or less the core-tag-set (i.e. the elements that are available in all TEI-documents). It is indeed a very simple DTD and every piece of "SGML conformant software" is able to process it. A drawback of TEI-Lite is that the base tag-sets (e.g. the base tag-set for the transcription of speech) and the additional tag-sets are not available[3].

3. 3. TEI IS NOT XML

XML defines a subset of SGML and restricts SGML in a number of ways[4].

The TEI P3 DTD uses some of these features. The most striking difference is the prohibition of tag omission and quasi-prohibition of short-tags. The TEI *interchange format* doesn't allow tag omission either, but since the same DTD is used for both *local processing*, which allows tag omission, and document *interchange*, the final document is not a valid XML-document. Similarly the TEI interchange format prohibits the use of *SHORTTAG* concerning the element name which is allowed for local processing. The omission of attribute names is allowed in both TEI-formats, but not allowed in

XML. Further differences concern the inclusion of subdocuments, the declaration of obligatory elements in free order (with '&'), and the case-insensitive use of element- and attribute names. It is possible to divide the differences between the SGML-subset used by the TEI-DTDs and the subset used by XML into two groups.
XML restricts SGML in such a way that the TEI-DTD can't be used
The SGML declaration used by XML is more general in some points than the SGML declaration employed by TEI P3

The first point concerns the expression language (e.g. the prohibition of the grouping operator '&'). An example of the second point are the capacities of XML, though, within TEI format, it is not prohibited to increase the predefined capacity-information.

4. PROJECT

The project described aims at combining the advantages of the TEI-approach to the transcription of speech with the advantages of XML and vice versa.

4.1. DOMAIN

A Special Research Group has the goal of a robot as an 'artificial communicator', that has to solve a simple assembly task. The robot gets instructions from humans. One aim of the project, therefore, is the acoustic perception of the spoken word. Using a „Wizard of Oz Scenario" a group of linguists conducted a study to analyze the language of humans who believe that they are speaking with a robot. In total, they examined about 50 persons and they collected more then twenty hours of speech data. These data were transcribed as shown in the example[5]:

Header:
Versuch: 09[↵]
Instrukteur (HO) m, Konstrukteur (K) 6,
Versuchsleiter (V) 1[↵]
Dialogart: Simulierte Mensch-Maschine-
Kommunikation[↵]
Bauart: Flugzeug[↵]
Dialogdauer: 53 min[↵]

Content(extracted):

```
09HO021<hum: atmen> verbinde <--> die
<--> Stange mit sieben L"ochern <sil:
1> <hum: schlucken> <hum: atmen> mit <-
> der <-> {orangenen}
<attrib: z"ogern> eckigen Schraube
<hum: atmen> und dem gelben W"urfel im
letzten Loch <--> der <-->
f"unfl"ochrigen Stange <sil: 1> <hum:
atmen> und im mittleren Loch der <->
Stange mit sieben L"ochern. [↵]
K<sil: 3> ich habe verstanden. <sil: 3>
Ihre Anweisung wird bearbeitet. <hum>
<sil: 8> einen Augenblick bitte. <sil:
12> der Arbeitsschritt ist beendet.
bitte fahren Sie fort. </hum: atmen>[↵]
```

The transcriptions were annotated in a specially developed encoding scheme [5] and are available as ASCII-Texts and as MS-Word documents. Furthermore, the original dialogs were recorded and stored in audio files (in raw - format). My aim was to transform the transcripts into an XML notation and to incorporate a link to their respective audio-recordings.

4. 2. TARGET FORMAT

I developed a document type definition (DTD) which uses all restrictions of XML on the base of the TEI-Guidelines. The document instance is usable both as an XML-instance and as an TEI-document[6]. The only thing that needs to be changed in the final document is the pointer to the DTD (i.e. <!DOCTYPE TEI.2 system "XMLTEI.dtd"> vs. <!DOCTYPE TEI.2 public "TEI ..." [...]>) .The TEI-annotated version of the extract shown is given in the next example:

Header:

```
<title> Versuch:  09 </title>

<particdesc>
<person
id=HO><p>Instrukteur<p></person>
<person
id=K><p>Konstrukteur<p></person>
<person
id=V><p>Versuchsleiter<p></person>
</particdesc>
```

Content(extracted):

```
<u who="HO" n="021" id=refno93
idref=refno93>
<address>&sig021;</address>
<div><vocal desc="atmen"> verbinde
<pause type="long"> die
<pause type="long"> Stange mit sieben
L&ouml;chern <pause dur="1">
<vocal desc="schlucken"> <vocal
desc="atmen"> mit <pause type="short">
der <pause type="short"> <shift
feature="voice" new="z&ouml;gern">
orangenen <shift feature="voice"
new="normal"> eckigen Schraube <vocal
desc="atmen"> und dem gelben
W&uuml;rfel im letzten Loch <pause
type="long"> der
<pause type="long">
f&uuml;nfl&ouml;chrigen Stange <pause
dur="1"> <vocal desc="atmen"> und im
mittleren Loch der <pause type="short">
Stange mit sieben
L&ouml;chern.</div></u>
<u who="K" id=refno21
idref=refno21><div><pause dur="3"> ich
habe verstanden. <pause dur="3"> Ihre
Anweisung wird bearbeitet.
```

```
<anchor id=xno149> <pause dur="8">
einen Augenblick bitte. <pause
dur="12"> der Arbeitsschritt ist
beendet. bitte fahren Sie fort.
<anchor id=yno149><kinesic
start=xno149 end=yno149
desc="atmen"></div></u>
```

The entities `&sigXXX;` are newly introduced and refer to the audio-files.

4. 3. METHOD

The method I employed in constructing an implementation that meets the needs of TEI as well as the requirements of XML is summarized as follows:

analysis of the used encoding scheme;
mapping of the used annotations on the tag-set TEI P3;
extending of the TEI-DTD in the manner described in chapter 29 of the TEI-Guidelines if necessary;
annotation of the sample text (This annotation uses all phenomena of the original encoding scheme within the TEI-framework.);
writing of a DTD within the restrictions of XML to define the document built in the previous step (step 4).

I want to stress that this method is also appropriate for the transformation of TEI-annotated documents. In case the source-document is indeed already a TEI-instance, one skips step 1-4 and "starts" at point 5.

4. 4. DEFINING THE XML-BASED TEI-CONFORM DTD

The requirements of the target document instance(s) are twofold: the SGML declaration of XML is both more restrictive and more liberal than the SGML declaration of TEI P3. This leads to a target-annotation that must meet a core declaration, i.e. a very restrictive declaration allowing just those extensions which are used by both declarations. Fortunately such an SGML-declaration exists. Because XML already uses a SGML-declaration, that is quite strict, the new SGML-declaration is very similar to it. The main change concerns XML's definitions of the *capacities*[7]. The SGML declaration of TEI P3 differs from it mainly in the use of the SUBDOC, OMITTAG, and in a case-sensitive treatment of element- and attribute names.

4. 5. TRANSFORMATION OF THE TRANSCRIPTION

After developing the DTD, the only task remaining is transforming the transcriptions. Since the encoding scheme of the transcriptions that I mentioned, were formally defined in an BNF (i.e. a formalism describing the grammar of the notation, called Backus-Naur Form), this step was processed automatically with a *Perl*-program.

4.6. PRESENTATION

One of the strengths of an XML-based preparation of transcriptions as well as of other texts is the possibility of creating an elegant document presentation. At the time writing no genuine XML browser is available yet. Nevertheless, a presentation of XML documents with an SGML browser is possible[8]. A presentation using frames was developed. In total, four frames were displayed. Every frame shows the same transcription differently. The meta-information about the utterances, for instance, is displayed in one frame in German and in an other frame in English. Furthermore, one frame displays only the number of the utterance, the associated speaker, and two icons. One click on the one icon causes the originally recorded sound to be played and one click on the other icon causes a window to open, which displays the transcription without annotations.

5. CONCLUSION AND REMAINING WORK

The paper describes a way of transforming TEI documents into documents which are both TEI-documents and meet the requirements of XML. I want to stress once more, that this approach is not limited to the base tag-set "Transcription of Speech". One desirable aim would be to formalize the steps mentioned in a way that would make possible an automatic generation of the XML-DTD. Current research by Henning Lobin at the University of Bielefeld is concerned with the unification of different SGML-documents, and could lead to precisely such a formalization.

NOTES

[1]The relation of XML to the other standards had been described as: XML is a kind of SGML-- rather than HTML++.

[2]I remember working on a book which had to be edited. The book contained (among other things) four papers of socio-linguists who worked at the same university. These four researchers used three different notations for their transcriptions.

[3]If these tag-sets are used in conjunction with the TEI-Lite, one faces the same problems with the TEI P3 as described above.

[4]At the time of the writing of this paper the XML recommendation defines valid documents that are not valid SGML documents. This will change in the final specification either through a change of SGML or through a change of XML.

[5] [↵] indicates a carriage-return.

[6]I did not use the XML syntax for empty elements. This general problem will be solved soon (cf. Footnote 8)

[7]Even this is not necessary because – as stated above – the capacity information in the declaration of the TEI SGML-declaration may be increased.

[8]The *existing SGML browsers restrict SGML* in a way that resembles SGML restrictions of XML. Not surprisingly, the book "SGML on the Web" by Y. Rubinsky and M. Maloney is widely regarded as a fundamental source for describing the concepts of XML.[7]

REFERENCES:

1 Bray, T., Paoli, J., Sperberg-McQueen, C. M. (1998). Extensible Markup Language (XML). W3C Recommendation 10-February-1998

2. Burnard, L. and Sperberg-McQueen C.M. (1995). TEI Lite: An Introduction to Text Encoding for Interchange, TEI U5 June 1995.

3. Burger, S. (1997), Transliteration spontansprachlicher Daten - Lexikon der *Transliterationskonventionen* - VERBMOBIL II, Technisches Dokument 56 April 1997. Universität München.

4. ISO 8879 ISO (International Organization for Standardization) Information processing -- Text and Office Systems -- Standard Generalized Markup Language (SGML). First edition -- 1986-10-15. International Organization for Standardization. Geneva.

5. Fink, Johanntokrax, and Schaffranietz (1995). A Flexible Formal Language for the Orthographic Transcription of Spontaneous Spoken Dialogues, Proc. European Conference on Speech Communication and Technology. Madrid.

6. Kohler, K., Lex, G., Pätzold, M., Scheffers, M., Simpson, A., and Thon, W. (1994). Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL, Technisches Dokument Nr. 11 September 1994, IPDS Kiel.

7. Rubinsky, Y., and Maloney, M. (1997), SGML on the Web : small steps beyond HTML - Prentice Hall. (Upper Saddle River, N.J.)

8. Sperberg-McQueen, C. M. and Burnard, L., editors (1994). Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative. Chicago, Oxford.

9. Witt, A., Lüngen, H. and Gibbon, D. (1997), Standardisierung orthographischer Transkriptionen: Ein SGML/TEI-basierter Vorschlag für VERBMOBIL. Memo Nr. 117 January 1997. Universität Bielefeld.