Embedding IMDI metadata into a large phonetic corpus

Oliver Schonefeld*, Jan-Torsten Milde†

*Technical Faculty
Bielefeld University, Bielefeld, Germany
oschonef@TechFak.Uni-Bielefeld.de

†Computer Science Department
University of Applied Sciences, Fulda, Germany
Jan-Torsten.Milde@fh-fulda.de

Abstract

The paper shows the set up of a large phonetic corpus (the LeaP corpus), how its metadata is structured and transformed into an extended IMDI/ISLE metadata structure, how this structure has been transcoded into the TASX metadata format and finally has been intergrated into the LeaP corpus.

1. Introduction

The paper shows the set up of a large phonetic corpus (the LeaP corpus), how its metadata is structured and transformed into an extended IMDI/ISLE metadata structure (Wittenburg et al., 2000), how this structure has been transcoded into the TASX metadata format and finally has been intergrated into the LeaP corpus.

The LeaP project (Milde and Gut, 2002a) explores the acquisition of prosody by second language learners of both German and English. It focuses on three areas of prosody: stress assignment on both the word and the phrase level, sentence intonation and speech rhythm. In addition, the form and function of gestures in non-native speech have been analysed.

In a period of two years a large set of recordings ($>$ 300) of second language learners' speech have been made and phonologically annotated. From this data an TASX-annotated spoken language corpus has been set up. A single recording contains about 3000 tags distributed on up to 8 levels of annotation. As a result, the LeaP corpus contains almost 1 million tags.

When working with a corpus of this size and structure it becomes necessary to integrate non linguistic metadata into the corpus. Metadata can be used to create sub corpora (e.g. a sub corpus only containing native German speakers) and even more important to optimize corpus queries.

A large quantaty of metadata has been collected for the LeaP corpus. This includes metadata describing the experimental setup and the recording conditions, metadata describing the speaker in some detail and metadata storing information on motivations and attitudes of the speaker:

1. *experimental setup*: date of the recording, location of the recording, interviewer and target language, information about the transcription, technical information

2. *recorded speaker*: age, sex, native language, foreign language competence, age of first contact with target language, mode of contact (formal language training or unsupervised learning) duration and mode of stays abroad in target language speaking countries, duration and mode of formal prosody training (if available), existing knowledge about prosody of the target language

3. *motivations and attitudes*: rate the importance of sounding like a native speaker (on a scale from 1 to 5), reasons for learning the target language, rate the importance of native like pronounciation vs. other language related abilities (knowledge of grammar, size of active vocabulary etc.), interest in music (on a scale from 1 to 5), musical experience (on a scale from 1 to 5), musical competence (on a scale from 1 to 5)

The metadata has first been collected using a simple Java based metadata editor, later the metadata were translated to IMDI and then transcoded to fit into the corpus format. The next section is going to describe the underlying TASX format in more detail. Section 3 explains how metadata is included in the corpus. Finally a short summary is given.

2. The TASX format

All linguistic data in our system is stored in an XML-based format called TASX: the *T*ime *A*ligned *S*ignal data e*X*change format. With TASX it becomes possible to to transform, query and distribute the content of multimodal corpora, and to perform adequate linguistic analysis (Milde and Gut, 2002b).

A TASX-annotated corpus consists of a set of *sessions*. A session basically represents a single *recording* or *experiment* of a multimodal corpus. Each session is holding an arbitrary number of descriptive tiers, called *layers*. The layers of a session are treated as time aligned sets, which are directly or indirectly linked to the primary data. Layers are used to distribute the annotations of a session and thus simplifying the information access. Each layer consists of a set of independent *events*. The events store textual information (e.g. a syllable or a handform) and are linked to the primary audio data by two time stamps denoting the interval of this event. Events form the annotations atoms of a TASX annotated corpus. With respect to the formal definition of TASX, events are the leaves of the XML tree. No further formal restriction is imposed on the content of an event.

Sessions, layers and events all carry linking attributes. All elements of a TASX annotated corpus must provide a uniqe ID (*s-id*, *l-id* and *e-id*)). In addition cross-references between the elements can be specified using the optional (*ref*) attribute. A validating XML parser should be able to check the existence of the IDs and, if cross references are provided, the consistincy of the intra document linking. The ID/IREFS mechanism represents a means to further restrict the *content* of a TASX annotated corpus, e.g. a session *ref*-attribute could register all IDs of the enclosed layers. The following DTD fragment formalizes the TASX format:

```
<!-- corpus data -->
<!ELEMENT tasx (meta*,session+)>

<!ELEMENT session (meta*,layer+)>
<!ELEMENT layer (meta*,event+)>
<!ELEMENT event (#PCDATA,meta*)>

<!-- atributes -->
<!ATTLIST session
        s-id ID #REQUIRED
day CDATA #REQUIRED
        ref IDREFS #IMPLIED
        month CDATA #REQUIRED
        year CDATA #REQUIRED>

<!ATTLIST layer
       l-id ID #REQUIRED
       ref IDREFS #IMPLIED>

<!ATTLIST event
          e-id ID #REQUIRED
          start CDATA #REQUIRED
          end CDATA #REQUIRED
          ref IDREFS #IMPLIED
          mid CDATA #IMPLIED
          len CDATA #IMPLIED>
```

2.1. TASX-level 1

A TASX-annotated corpus, that *directly* links the primary data and the corpus data by defining a temporal interval in the start/end attributes of an event, is called a *TASX-level 1* corpus. TASX-level 1 provides a solution for one of the most common problems of XML-annotated multimodal corpora: the temporal *overlap* of annotation units. XML files define a tree structure and as such an overlap of opening- and closing tags is not allowed. This limitation is too restrictive for a large number of linguistic application areas. Within the TASX-annotated XML file the events are ordered linearly, but their scope is defined in the interval encoded in the attributes. Accordingly events may overlap in time. This mechanism can also be applied to primary data, which has no intrinsic temporal order, respectively to data, where the temporal order is lost (e.g. it is complex to reconstruct the temporal order of SyncWriter files, because the segmentation lists do not carry any temporal information). In order to describe temporal overlap in such primary data, a set of linearly ordered reference points has to be defined. This can be achieved by simply refering to the set of natural numbers.

2.2. TASX-level 2

A TASX corpus, which extends the direct temporal linking of events to *linking events to other events* is called a *TASX-level 2* corpus. TASX-level 2 allows to define hierarchical relations between annotation layers. These inter layer relations can be established, because the formal description of TASX does not restrict the values of the start/end attributes in any way. Therefore arbitrary strings can be used to describe the relations to other layers and events on these layers. These strings might contain XPointer or XPath expressions (Wilde and Lowe, 2002) or being formulated in any other suitable syntax. With respect to linguistic research, this TASX approach allows to represent cross level relations, e.g to denote the hierarchical relation between words and syllables. In a case study in conjunction with Voormann (Voormann et al., May 2004) hierarchical relations between different layers of the LeaP corpus have been computationally constructed. Implementations in related tools follow the same approach, e.g. the Elan annotation tool designed by Brugman und Wittenburg (Brugman and Wittenburg, 2001) provides a mechanism to constrain the relations between annotation layers.

TASX-level 1 and 2 are comparable to standard approaches currently discussed in the field of computational corpus linugistic (Bird and Liberman, 1999). Schmidt provides Exmaralda, a tool for conversational analysis, which follows the Bird & Liberman approach (Schmidt, 2001). Alternatively stand-off markup is proposed by MATE (Dybkjaer et al., June 1999), respectivly NITE (Carletta et al., 2002). , while Kipp (Kipp, 2001) specifies hierarchical relations between tiers.

2.3. TASX-level 3

TASX defines a *data centric* information model. It focuses on the organisation of the data and leaves the internal logical structure of the primary data untouched. The TASX model basically organizes the data of a session as a two dimensional array. A row of the array is equivalent to an annotation layer, each field of the array is mapped to an event. The content of the fields is unrestricted.

The clear advantage of such an approach is the fact, that arbitrary event descriptions can be gathered in a TASX conformant corpus. Within in the context of the LeaP corpus we were able to encode orthographical representations of words and syllables, but also phonemes encoded as SAMPA strings and phrasal tone gradients encoded according to the ToBI standard (Beckman and Elam., 1994).

It is also possible to include primary data with a more complex structure, as long as the data can be serialized and represented as a linear string. One possible solution is to encode the data as Base64 ASCII strings. This approach is used inside the TASX-annotator to store multi-line comments entered by the user. Finally XML-annotated strings could be stored as the content of an event. In this case, the special characters of XML have to be encoded with their entity counterpart.

It is due to this flexibility, that TASX is able to integrate a large number of currently available representational formats for linguistic data without information loss. At the same time this also marks a central problem of the data cen-

tric approach. The internal structure of the data stored in an event is completly transparent to TASX. As a consequece, there is no direct way to ensure the consistency and validity of the data stored (at least not with the formal definition of TASX being descriobed as a DTD).

As such TASX seemingly dismisses the two most important advantages of XML: the *structure guided creation* of annotated content, based on the formal description in form of a DTD and the *automatic verification of the linguistic structure* of a corpus using a standard validating XML parser.

With *TASX-level 3* we try to define a balanced compromise between the data centric and the structure centric approach. The approach tries to decouple the XML elements of the underlying data centric TASX format from the XML elements describing the semi structured linugistic data. To achieve this, a TASX namespace has been defined. While this is not possible with DTDs, we used XML schema to formally define the namespace. An XML parser processing a TASX corpus will be able to distinguish between the TASX elements and the embedded elements. The approach thus allows to encode tree like structures with XML and stores them as part of a TASX corpus. The embedded XML structures must be wellformed. TASX-level 3 therefore allows to setup corpora that e.g. combine syntactic trees and phonetic transcriptions. References between the different layer can be established.

Despite of its simplicity, the TASX-format is powerful enough to encode most of the corpus annotation formats currently in use. Indeed a number of format transformation programs have been implemented. In order to e.g. reconstruct the equivalent annotation graphs (Bird and Liberman, 1999) representation of a TASX annotated corpus, one only has to collect the time stamps encoded in the start and end attributes of the event tags, sort them and then produce the timeline. Finally the time stamps of the events have to be replaced by references to the timeline.

3. Integrating IMDI/ISLE metadata

Metadata can be assigned to all levels of the TASX format: to the complete corpus, each session, each layer and each event. On all levels, the metadata is stored as a vector of descriptions, each consisting of an attribute/value pair. A vector represents a separate metadata section. Each of the metadata sections is identified by a unique identifier (*m-id*).

It might be reasonable to extend the metadata description in a way that tree structured data can immediately be described by XML annotations. Currently we rather use a simpler version with linear structure.

The metadata section can be used to store linguistic and extra linguistic metadata as well as storing *tool-orientied* metadata. The metadata sections thus enable the exchange of configurations between otherwise incompatible linguistic tools. The following DTD fragment gives a formal definition of the meta element defined in TASX.

```
<!-- metadata -->
<!ELEMENT meta (desc*)>
<!ELEMENT desc (name,val)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT val (#PCDATA)>

<!ATTLIST meta
        m-id ID #REQUIRED
        ref IDREFS #IMPLIED
        access CDATA #IMPLIED
        level CDATA #IMPLIED>
```

The IMDI/ISLE standard defines a rich metadata element set for the use in language resources. Furthermore the standard uses an XML-based data format and a powerful metadata editor is provided by the Max Plank Institute, Nijmegen (IMDI-Team, July 2003), (IMDI-Team, August 2001). We therefore decided to use IMDI in the LeaP project.

Unfortenatly the IMDI/ISLE metadata set is not well suited for phonetic corpora. In the case of the bi-lingual/trilingual LeaP corpus, multiple languages have to be encoded in the metadata. This has not been possible with the chosen IMDI/ISLE metadata set. As a result we extended the element set by defining a number of new categories.

Another drawback of the IMDI/ISLE approach is its clear orientation towards field linguistics. This did not fit well into the experimental setup of the LeaP project.

Finally IMDI/ISLE uses repeated category names. While the IMDI hierarchy disambiguates the category names, it still poses some problems when formulating corpus queries with XQuery.

The IMDI/ISLE metadata set is formally described with an XML schema. The overall structure of IMDI is comparably more complex then the TASX approach. Still a lossless transformation to TASX was achieved. The overall metadata conversion process was executed in three steps:

1. plain text to XML conversion: a converter was implemented, to convert the plain text version of the metadata description into IMDI/ISLE conformant XML

2. mapping of hierarchically organized ISLE/IMDI onto a linear XML structered annotation: a transformation programm was implemented to convert IMDI/ISLE to TASX.

3. integration into the LeaP corpus: logically the LeaP corpus is a single file, while physically it is distributed on more then 600 files. The metadata files had to be integrated into the linking concept and the file structure

Step 1 of the integration process is needed to reintegrate legacy metadata into the final corpus. In the beginning of the LeaP project a simple metadata editor had been implemented. This system stored the metadata in plain text and therefore a conversion to XML was required. In addition a number of new metadata categories had to be included and old categories had to be mapped onto IMDI category names.

In Step 2 the hierarchical structure of IMDI is replaced by its linear TASX representation. This process is performed in a fully automatic way using an XSL-T program (`xml2xpath.xsl`). The program traverses the XML tree structure in a topdown, depth-fist manner. For each leaf, the program generates the corresponding XPath expression.

```
<meta id="IMDI/Session">
<meta id="IMDI/Session/MDGroup">
<meta id="IMDI/Session/MDGroup/Location">
<meta id="IMDI/Session/MDGroup/Content">
<meta id="IMDI/Session/MDGroup/Content/CommunicationContext">
<meta id="IMDI/Session/MDGroup/Content/Genre">
<meta id="IMDI/Session/MDGroup/Content/Languages/Language">
<meta id="IMDI/Session/MDGroup/Participants/Participant1">
<meta id="IMDI/Session/MDGroup/Participants/Participant1/Languages/Language1">
<meta id="IMDI/Session/MDGroup/Participants/Participant1/Languages/Language2">
<meta id="IMDI/Session/MDGroup/Participants/Participant1/Languages/Language3">
<meta id="IMDI/Session/MDGroup/Participants/Participant1/Languages/Language4">
<meta id="IMDI/Session/MDGroup/Participants/Participant1/Keys">
<meta id="IMDI/Session/MDGroup/Participants/Participant2">
<meta id="IMDI/Session/Resources/AnnotationUnit">
```

Figure 1: Example structure of the transcoded IMDI metadata.

This includes the position of each node/attribute on each level. The XPath expression is then stored as the name of the TASX metadata element, while the content of the leaf is stored in the val element (see figure 1). The process fully conserves the tree structure of the input file. In addition querying the original XML file becomes much easier, as the generated XPath expressions can be applied to the IMDI file directly. The transformation works on arbitrary XML files, thus making it possible to integrate arbitrary XML structures into TASX annotated corpora [1].

Step 3 of the conversion process integrates the generated TASX metadata sections into the LeaP corpus. XML does not provide means to modularize larger files. We therefore used the entity mechnism of DTDs to integrate the physically distributed parts of the corpus. So instead of merging all partial XML files into one large corpus file, a list of entities is generated. The program (`generateLinkedCorpus.pl`) traverses the file system and searches for XML files. The separate XML files either contain a TASX session or a metadata section. For each file found an entitiy declaration is defined. Only relative file paths are used, allowing to move the corpus without creating stale file references. The entity declarations are placed in the declaration subset of the central corpus file. Finally each entity is listed and the whole corpus is enclosed with the <`tasx`> root tag.

4. Conclusion

This paper showed how to set up a large phonetic corpus using the TASX approach. The TASX format is well suited to encode most of the corpus currently in use. As such it becomes possible to exchange linguistic data between otherwise incompatible tools. The IMDI metadata approach fits very well into the TASX based approach developed for the LeaP project. As such IMDI tools can now be used to create metadata resources which are fully compatible with TASX.

5. References

Mary E. Beckman and Gayle Ayers Elam. 1994. Guidelines for ToBI Labelling. Technical report, Ohio State University. Version 3.0, March 1997.

S. Bird and M. Liberman. 1999. A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Science, University of Pennsylvania.

Hennie Brugman and Peter Wittenburg. 2001. Mpi tools for linguistic annotation. In Peter Buneman Steven Bird and Mark Liberman, editors, *IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA*.

J. Carletta, D. McKelvie, and Isard A. 2002. Supporting linguistic annotation using xml and stylesheets. In G. Sampson and D. McCarthy, editors, *Readings in Corpus Linguistic, Continuum International*.

L. Dybkjaer, M. B. Moeller, N. O. Bernsen, J. Carletta, A. Isard, M. Klein, D. McKelvie, and A. Mengel. June 1999. The mate workbench. In David Traum, editor, *Proceedings of ACL'99, Demonstration Abstracts. University of Maryland*, pages 12 – 13.

IMDI-Team. August 2001. Vocabulary Taxonomy and Structure, Version 1.0. Technical report, MPI Nijmegen.

IMDI-Team. July 2003. IMDI Metadata Elements for Session Descriptions, Version 3.0.3. Technical report, MPI Nijmegen.

Michael Kipp. 2001. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of the Eurospeech 2001, Aalborg*, pages 1367 – 1370.

J.-T. Milde and U. B. Gut. 2002a. A prosodic corpus of non-native speech. In B. Bel and I. Marlien, editors, *Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002. Aix-en-Provence: Laboratoire Parole et Langage*, pages 503 – 506.

J.-T. Milde and U. B. Gut. 2002b. The tasx-environment: an xml-based toolset for time aligned speech corpora. In *Proceedings of the third international conference on language resources and evaluation (LREC 2002, Gran Canaria*.

T. Schmidt. 2001. Gesprächstranskription auf dem Computer - das System EXMARaLDA. *Gesprächsforschung, http://www.gespraechsforschung-ozs.de*, 2.

Holger Voormann, Ulrich Heid, Jan-Torsten Milde, Ulrike Gut, Katrin Erk, and Sebastian Pado. May 2004. Flexible querying of xml-encoded multi-layer corpora. In *Proceedings of LREC 2004, Lisbon*.

Erik Wilde and David Lowe. 2002. *XPath, XLink, XPointer, and XML*. Addison-Wesley Professional.

P. Wittenburg, D. Broeder, and B. Sloman. 2000. Eagles/isle: A proposal for a meta description standard for language resources, white paper. In *LREC 2000 WS, Athen*.

[1] The TASX-annotator is able to process these strings and provides an GUI interface to manipulate TASX metadata (Milde and Gut, 2002b).