Sustainability of Linguistic Resources

Stefanie Dipper[†], Erhard Hinrichs[‡], Thomas Schmidt[*], Andreas Wagner[**], Andreas Witt[‡]

[†]Universität Potsdam
dipper@ling.uni-potsdam.de
[‡]Eberhard-Karls-Universität Tübingen
eh@sfs.uni-tuebingen.de
andreas.witt@uni-tuebingen.de
[*]Universität Hamburg
thomas.schmidt@uni-hamburg.de
[**]Universität Duisburg-Essen
andreas.wagner@uni-due.de

Abstract

This paper describes a new research initiative addressing the issue of sustainability of linguistic resources. This initiative is a cooperation between three linguistic collaborative research centres in Germany, which comprise more than 40 individual research projects altogether. These projects are involved in creating manifold language resources, especially corpora, tailored to their particular needs. The aim of the project described here is to ensure an effective and sustainable access of these data by third-party researchers beyond the termination of these projects. This goal involves a number of measures, such as the definition of a common data format to completely capture the heterogeneous information encoded in the individual corpora, the development of user-friendly and sustainably usable tools for processing (e.g. querying) the data, and the specification of common inventories of metadata and terminology. Moreover, the project aims at formulating general rules of best practice for creating, accessing, and archiving linguistic resources.

1. Introduction

This paper describes a new DFG[1] funded project (10/2005 – 12/2008) on preparation of language resources for assuring an accessible dissemination and a sustainable storing of these corpora. A main aim of the project is a practical one: resources acquired in long-term projects from three 'Collaborative Research Centres' have to be converted in one or several formats to be sustainably usable by researchers and applications. Furthermore it is envisaged to provide a unified access for the heterogeneous data acquired in the different involved projects. In addition to the preparation of already existing language corpora, general methodologies and 'Rules of Best Practice' should be developed.

The paper is structured as follows: Section 2. describes the resources of the three Collaborative Research Centres. These Centres are the SFB 538 'Multilingualism' at the University of Hamburg, the SFB 632 'Information Structure' at the University of Potsdam and the Humboldt University Berlin, and the SFB 441 'Linguistic Data Structures' at the Eberhard Karls University Tübingen.

Section 3. describes the technical aspects of the project. Especially the aspect of a data format usable as a general or meta-format for the formats in the three SFBs is adressed. Because of the heterogeneity of the formats we expect that such a format could serve as a meta-format for a wide range of XML-based annotation schemes.

Section 4. addresses the use of an appropriate set of meta data and the integration of formally defined termologies for enhancing interoperability of the annotated data.

The paper ends with some remarks on the 'Rules of Best Practice', copyright issues and rights of personality.

2. Annotated resources and annotation schemes

2.1. Hamburg

2.1.1. Annotated resources

The research centre on multilingualism at the University of Hamburg comprises 14 projects doing research on diverse aspects of multilingualism. All projects work empirically, basing their analyses on digital corpora of written or transcribed spoken language. Apart from the spoken/written distinction, these data differ with respect to many more dimensions, but very roughly fall in one of the following categories:

- Longitudinal first language acquisition data of bilingual children - these are mostly transcriptions of video recordings of child/caretaker interactions;
- Other language acquisition data - this comprises sampled (as opposed to longitudinal) L1 acquisition data of mono- or bilingual children as well as data from L2 learners and from children with specific language impairments;
- Multilingual spoken communication data - this includes, for instance, transcribed radio broadcasts of Inter-Scandinavian communication, transcriptions of interpreter-mediated doctor/patient communication, Japanese/German expert discourse (e.g. business or academic communication) and semi-structured interviews with bilingual speakers from the Faroe Islands;

[1]Deutsche Forschungsgemeinschaft, i.e. the German Research Foundation.

- Historical texts - examples of these are Old Swedish and Old Danish bible translations, 19th century letters by Irish emigrants and Old French legal documents;
- Modern texts - this comprises a parallel corpus of English and German business texts as well as a parallel corpus of popular science writing.

Apart from this conceptual diversity, the data in their original form also exhibited a great diversity on the technical level, in particular with respect to their storage formats (ranging from RDB-like over text-based to binary formats) and the tools with which they could be created, edited and analysed. Since this diversity made data exchange and data reuse extremely difficult, the EXMARaLDA system was developed to give these data a common structural backbone and thus to facilitate data exchange and data reuse as well as the construction of multi-purpose transcription and query tools.

2.1.2. Annotation schemes

Building on the idea of the annotation graph framework (Bird and Liberman, 2001), EXMARaLDA uses a time-based data model. This means that the primary relation between any two entities in a data set is established via their reference to a timeline, and not via their position in some other structure like, for instance, an ordered hierarchy. All non-temporal relations, like hierarchical inclusion or entity/feature relations, are regarded as secondary features that can be derived from this temporal structure. EXMARaLDA defines a *basic* and an *extended* data model for working with linguistic data.

The basic data model (a "Basic-Transcription") is a variant of the "Single Timeline, Multiple Tiers" model which is also used by a number of other systems or tools like Praat, ELAN, the TASX annotator or ANVIL. In general, these kinds of data model organise individual descriptions (*events*) into a number of *tiers* (or layers) and relate them to one another by assigning each description a start and an end point from a single, fully ordered *timeline*. In addition to that, the basic data model in EXMARaLDA requires that, firstly, no two events within a tier must overlap. Secondly, each tier can be assigned a *speaker* and must be assigned a *category*. Categories, in turn, fall into three *types*: T(ranscription) for tiers in which verbal behaviour is described, D(escription) for tiers in which non-verbal behaviour is described, and A(nnotation) for tiers in which stretches of transcribed speech are further categorised. This data model has proven adequate for the process of data creation as well as for many data visualisation tasks. In particular, its theory-neutrality makes it applicable for a wide range of researchers, its comparative simplicity facilitates the construction of intuitive user interfaces, and its similarity to the models of other systems (mentioned above) makes data exchange between EXMARaLDA and these systems a fairly straightforward matter.

The extended data model (a "Segmented-Transcription") caters for more complex tasks like querying and extensively annotating data, as well as for additional types of visualisation and for long term archiving. On top of the temporal structure encoded in the basic data model, it allows for the representation of additional linguistic structure. Most importantly, this means a segmentation of transcribed speech events into words and entities like turns, utterances or intonation units. Since these linguistically motivated units and the temporally motivated units in a Basic-Transcription do not have a uniform relation to one another (i.e. neither do their boundaries coincide in a regular way nor is one systematically included in the other), encoding them both in one document requires additional structural complexity. This is attained by allowing for a *bifurcating*, partially ordered timeline instead of the fully ordered one in the basic data model. In practice, the additional linguistic structure in a Segmented-Transcription is calculated automatically from the transcription convention regularities used in describing the temporal structure. Since different transcription conventions exhibit different such regularities, (and because they also define different linguistic units to begin with), data expressed in the extended data model is more dependent on specific linguistic theories than data expressed in the basic data model. For a more extensive discussion of EXMARaLDA's data model, see (Schmidt, 2005a) and (Schmidt, 2005b).

2.2. Potsdam / Berlin

The research centre SFB 632 at Potsdam University and Humboldt University Berlin investigates various facets of Information Structure (IS). IS concerns the means exploited by the speaker to structure discourse and utterances in order to convey information in a way that is optimised for the hearer in the given context. Languages differ a lot with regard to the means to express IS: by intonation, particles, word order, etc. The exact nature and interplay of many of these factors, however, is yet to be determined.

2.2.1. Annotated resources

The SFB consists of 13 individual research projects from disciplines such as theoretical linguistics, psycholinguistics, first and second language acquisition, typology, and historical linguistics. Following the overarching objective of providing a clearer picture of information structure, several of these projects are involved in collecting and analysing empirical data: Two projects examine the phenomenon of focus in different Western African languages; both carry out field studies for collecting data, which later is being annotated. One project investigates the role of IS in diachronic change, based on manuscripts of Old High German and Old English. Another project is developing a typology of the means for expressing IS. To this end, they have developed a language-independent questionnaire that is used to collect language data relevant for IS from speakers of typologically diverse languages, such as Hungarian, Greek, Georgian, Prinmi, Niue, Teribe, and Yucatec Maya; see, e.g., (Götze et al., to appear). Data sets elicited by the questionnaire consist of question-answer pairs, map-task dialogues, and short scenario descriptions. Finally, two projects focus on rhetorical and co-reference relations to address the relationship between discourse structure and IS.

According to the specific research interests of the indi-

vidual projects, this data is annotated at different levels, according to SFB-wide common guidelines. Diachronic data is annotated by morpho-syntactic features and givenness information; the Old High German translation of Tatian is furthermore word-aligned to the Latin source text. Typological data is annotated by phonetic/phonological information (breaks, pitch-range, tones, etc.), morpheme-to-morpheme translations, part of speech, syntactic constituents and their thematic roles, animacy, etc. Discourse-related data is enriched by annotations according to Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), co-reference and syntax annotations.

Currently, the corpora of these projects consist of several hundreds of data sets (for each of the languages of the typological data) and 20,000 German sentences (for discourse-related data).

2.2.2. Annotation scheme

To promote the active exchange of research hypotheses, the data is being collected in a single, uniform database, ANNIS. The database has to deal with highly heterogeneous data: First, primary data itself is heterogeneous, differing with respect to size (e.g., single sentences vs. entire articles), modality (monologue vs. dialogue), and language. Second, the annotations require data structures of various types (attribute-value pairs, trees, pointers, etc.). And finally, data is annotated by means of different, task-specific annotation tools: phonological, morphological and IS-related information, such as givenness, is annotated by EXMARaLDA, syntax by annotate, discourse structure by the RST Tool, and co-reference by MMAX2.[2]

Prior to import into the database, the data is mapped to a generic interchange format, PAULA[3]. This allows us to represent data annotations from different sources in a homogeneous way.

In our context, segments that annotations are attached to quite often overlap. The following example features an overlap between the phonemic and syntactic levels: at the phonemic level (= third tier), tokens 1 and 2, *de la* 'of the', are treated as one unit, whereas at the syntactic level, tokens 2–3, *la crème glacée* 'the ice-cream', form an NP constituent, cf. tier 4.

Token	*de*	*la*	*crème*	*glacée*
Gloss	some	the	cream	iced
Phonemic	**dla**		krEm	glase
Syntax	P	**NP**		

To account for such overlapping segments and for the heterogeneity of the data in general, PAULA uses an XML-based standoff architecture such that each annotation type is stored in a separate file. Annotations refer to the source text or to other annotations, by means of XLinks and XPointers. Building on proposals like in LAF (Linguistic Annotation Framework (Ide et al., 2003)), PAULA defines generic XML elements like `<mark>` (markable), `<feat>` (feature), `<struct>` (structure), and `<rel>` (relation), which allows us to represent, e.g., annotations attached to simple tokens as well as discontinuous segments, directed relations encoding anaphoric relations, and graphs to encode TIGER-like syntax trees or RST trees; for more details on the format, see (Dipper, 2005).

Currently, PAULA is used to represent data annotated by phonetic/phonological information, part of speech, morphology and lemma, syntax, rhetorical relations, anaphoric relations, and information structure.

For manual inspection of the data at multiple levels, we have developed the database ANNIS, (Dipper et al., 2004). ANNIS supports the concurrent visualisation of different types of annotations. The discourse view gives an overview on the discourse, while the table view enables easy and interactive access to multilayer annotations. The tree view displays syntactic structures.

The query facility of ANNIS offers a rich set of search operators that apply to primary data and annotations. It supports the use of wildcards and operators like precedence and dominance. Complex queries can be formulated by means of negation, logical "&" and "|" ('or'). Query results are displayed with the matching data (text and/or annotations) highlighted.

2.3. Tübingen

The principal concern of the collaborative research centre SFB 441 at University of Tübingen are the empiric data structures which feed into linguistic theory building. In order to approach this general issue from a considerable variety of research perspectives, SFB 441 comprises a number of projects (currently 15) each of which investigates a particular linguistic phenomenon, either concerning general methodological issues, or with regard to a particular language or language family. The respective research interests range from syntactic structures (such as coordination) in German and English, local and temporal deictic expressions in Bosnian/Croatian/Serbian or Portuguese and Spanish, to semantic roles, case relations, and cross-clausal references in Tibetan, to mention just a few.

2.3.1. Annotated resources

As empirical basis for their research, many projects create electronically accessible collections of linguistic data and prepare them to fit their particular needs. In most cases, these collections are corpora. However, a couple of projects deal with data (e.g. lexical information) which are more adequately represented by an Entity-Relationship based data model and thus are implemented in relational databases.

All data collections built within SFB 441 projects are assembled in one repository called TUSNELDA[4]. Especially, the different corpora are integrated into a common XML-based environment of encoding, storage, and retrieval. This integration is particularly challenging due

[2] `http://www.rrz.uni-hamburg.de/exmaralda/`; `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html`; `http://www.wagsoft.com/RSTTool/`; `http://mmax.eml-research.de/`.

[3] *Potsdamer AUstauschformat für Linguistische Annotation*, Potsdam Interchange Format for Linguistic Annotation.

[4] *TUebinger Sammlung Nutzbarer Empirischer Linguistischer DAtenstrukturen*, Tübingen collection of reusable, empirical, linguistic data structures.

to the heterogeneity of the individual corpora, which differ with regard to the following aspects:

- languages (e.g. German, Russian, Portuguese, Tibetan,...)
- text types / data types (e.g. newspaper texts, diachronic texts, dialogues, treebanks, ...)
- categories of information covered by the annotation / annotation levels (e.g. layout, textual structure, morpho-syntax, syntax, ...)
- underlying linguistic theories

The size of the individual corpora ranges from about 10,000 (Spanish/Portuguese spoken dialogues) to approx. 200 million words (German newspaper texts, automatically chunk-parsed). (Wagner, 2005) provides an overview of the corpora built by the individual SFB 441 projects.

2.3.2. Annotation scheme

Despite the diversity of the corpora in TUSNELDA, they all share the same generic data model: hierarchical structures. It is most appropriate to encode the phenomena captured in the TUSNELDA corpora by means of nested hierarchies, augmented by occasional "secondary relations" between arbitrary nodes in these hierarchies. This distinguishes TUSNELDA fundamentally from corpora whose annotation is based on other data models such as timeline-based markup of speech corpora or multimodal corpora (see especially subsection 2.1.). Such corpora encode the exact temporal correspondence between events on parallel layers (e.g. the coincidence of events in speech and accompanying gesture or the overlap of utterances) whereas hierarchical aspects are secondary. In TUSNELDA, however, hierarchical information (e.g. textual or syntactic structures) is prevalent, while capturing the exact temporal coincidence of different events in general is not of primary relevance in the research conducted within SFB 441.

Consequently, the annotation scheme developed for TUSNELDA encodes information as embedded (rather than standoff) annotation, immediately modelling hierarchical structures by XML hierarchies. Essentially, this decision rests on two major considerations. Firstly, this procedure makes it possible to utilise standard XML-aware tools (such as XML editors, format conversion tools, XML databases, or query engines), which are optimised for processing hierarchical XML structures so that they are well suited for embedded annotation, while providing at best rudimentary support for standoff annotation. Secondly, embedded annotation indeed is sufficient for encoding the data captured by the TUSNELDA corpora. Standoff annotation would be necessary if the structures to be encoded formed overlapping hierarchies, which cannot be modelled within a single XML document. However, the structures primarily encoded in the TUSNELDA do not overlap but can be integrated into a single hierarchy. For example, whereas syntactic structures constitute sub-sentential hierarchies, text structures define super-sentential hierarchies. Hence, these structures can be captured straightforwardly within a single XML document structure. Concurrent hierarchical units occur only marginally and are not of primary importance. These units concern the (physical) layout structure of the annotated texts, e.g. page boundaries. Such boundaries are marked by milestone elements (e.g. <pb/> for a page break), which do not violate the well-formedness of the document.

The following example, taken from the Tibetan Corpus in TUSNELDA (Wagner and Zeisler, 2004), illustrates the annotation of syntactic constituent structure, argument structure, and cross-clausal reference within an embedded environment. Syntactic constituents are encoded by the elements <clause>, <ntNode> (non-terminal node), and <tok> (token); their catogories are specified by <clauseCat>, <ntNodeCat>, and <pos> elements, respectively. Additional descriptions concerning individual constituents may be encoded within <desc> elements. A special case of such a description is the specification of the argument structure of a verb token. Especially, the subcategorisation frame realised in the current clause is encoded as <realFrame>, where each complement is represented by a <realComplement>. In the example, the first complement is not overtly realised within the clause (status="*empty*"). However, it is implicitly given by the context, i.e. it corresponds to the first complement of the previous clause. This correspondence is modelled by a <ref> (reference) element including a pointer (target) to the corresponding complement.

```
<clause>
  <ntNode>
    <tok>
      <orth>khra·phru·gu</orth>
      <pos>NOM:anim∼pers</pos>
    </tok>
    <ntNodeCat>NP</ntNodeCat>
    <desc>
      <case>Abs</case>
    </desc>
  </ntNode>
  <tok id="v6">
    <orth n="2">med-tshug</orth>
    <pos>VFIN</pos>
    <desc>
      ...
      <realFrame>
        <realComplement id="v6c1"
                        status="empty">
          <role>POSS</role>
          <ref target="v5c1"> </ref>
        </realComplement>
        <realComplement id="v6c2">
          <role>EXST2</role>
        </realComplement>
      </realFrame>
    </desc>
  </tok>
  <clauseCat>simple</clauseCat>
</clause>
```

3. Technical aspects

The description of the three 'Collaborative Research Units' in Hamburg, Potsdam, and Tübingen demonstrates

the large variety of language data and research interests. Consequently, different annotation schemes are used in these projects. In a way, this is a common situation in nearly every annotation related project and several standard solutions, e.g. the use of XSLT-based conversions, exist for dealing with this problem. But since this variety is also due to the variety of the given original data, i.e. audio, video, already annotated text, and raw text, we have to deal with a more fundamental problem. The different annotation schemes are based on different basic annotation methodologies. While some of the projects, especially projects in the SFB "Multilingualism", are using a graph-based methodology, others, especially the projects in the SFB "Linguistic Data Structures", use embedded markup where several annotation levels are mapped on a single annotation layer.[5]

3.1. Development of data formats

The data formats of the diverse collections of linguistic data should be converted to a uniform data format. This format must conform to widely accepted public standards. Furthermore, the data format must be supported by a wide variety of – ideally non-proprietary – software. Consequently the standards XML and Unicode have been chosen as a starting point. But using these standards does not suffice for a sustainable representation and storing of the data. Indeed, most of the existing data already use these standards.

XML and Unicode can be regarded as a base level of annotation. Two other important aspects of data formats for linguistic annotation are the use of the appropriate tagsets or annotation vocabularies and the use of a suitable data model for corpus annotation.

In the recent years, several general corpus annotation standards have been developed, e.g. TEI (Sperberg-McQueen and Burnard, 1994) or XCES (Ide et al., 2000). But, since in concrete projects specialised annotation schemes are important, further developments became necessary. The ISO TC37/SC4 developed an infrastructure, the already mentioned "Linguistic Annotation Framework (LAF)", to allow for combining general-purpose annotation formats (a dump-format) with specific annotation schemes (Ide et al., 2003).

Moreover, LAF defines a user extensible set of Data Categories and a user extensible Data Category Registry allowing for linking a corpus-specific annotation to a generic format. We intent to follow the LAF approach by combining the existing annotations, a generic annotation format (see below) and a linguistic terminology or ontology. (see also subsection 4.2.)

One of the main tasks of the project will be the development and implementation of a generic annotation format, i.e. a data model for the existing language data. The model must be applicable for all the language data already annotated in the projects involved.

[5]In this distinction a layer (or tier) is a technical realisation of an annotation, e.g. a single XML-file or a named directed path in an annotation graph, whereas level refers to an abstract level of description, e.g. in linguistics the levels of morphology, syntax, or semantics. (Bayerl et al., 2003)

In linguistics, hierarchical annotations are essential for embedding syntactic information in a corpus. Consequently a large percentage of the corpus data, especially the TUSNELDA data (see subsection 2.3.), require a hierarchical data model.

Graph based annotations, on the other hand, are the predominant data model for transcriptions of audio and video data and are the base of the EXMARALDA format. (see subsection 2.1.) Consequently, also these annotations must be represented in the uniform data format.

The data represented in PAULA (see subsection 2.2.) combine characteristics of hierarchical annotations and graph-based models. This is a typical situation for linguistic data annotated according to the standoff methodology (see (Thompson and McKelvie, 1997), (McKelvie et al., 2001)).

The variety of the data formats is a common situation for projects dealing with linguistic resources. Finding a meta-format suitable for covering all the data formats of the involved projects is a major task for the sustainable representation of corpus data. What is needed is a data format suitable for hierarchically annotated corpora as well as for graph based annotations.

As a starting point for such a format, we are currently evaluating the Nite Object Model (NOM, see (Carletta et al., 2003)).

3.2. Development of methods and tools for data distribution and data access

It is intended to produce and generate several distributions of language data. These distributions are optimised for distributing a whole collection of data, for a sustainable storing and for querying the data. The following methods of distribution are planned:

1. A human readable hardcopy of all corpus data;
2. An electronic version distributed as an offline medium (e.g. DVD);
3. A query interface accessible via the Internet.

For generating a printed version of the corpora XSLT stylesheets will be developed. The generated printable versions of the corpora can be archived and offered by libraries. For the electronic distributions (points 2 and 3) tools are to be implemented for the linguistic search in the data.

As described in section 2., in the SFBs involved query mechanisms for the respective data collections are already realised. However, the SFB's query mechanisms do have another focus, namely: power (the possibility of specifying complex search criteria), efficiency (short response time), and ease of use (input interfaces and output formats, should be comprehensible for linguists without advanced technical knowledge). Unfortunately, the criterion of sustainability is quite often in opposition to these criteria. For this reason new query mechanisms will be developed. For achieving a sustainable query interface, new tools will be based on XSLT and XQuery, since we expect these standards to be supported by software for a relatively long time.

4. Data Integretion

For an accessible storing of language corpora, the corpora must contain additional information. This additional information can be subdivided into two classes: (1) Information on the corpus itself, e.g. information on the participants of a conversation, the languages, the names of the transcribers, and (2) information on the meaning of the annotations, e.g. the tag w is used for annotating a word. The first class is traditionally termed "metadata". The second class of additional information is traditionally provided with the help of tag set documentations. At the moment, however, there is a tendency to use more or less formal apparata for this, namely terminologies or ontologies.

4.1. Metadata

It is intended to compile a comprehensive set of metadata. This set must adequately describe all the corpora of the SFBs. This implies that all the metadata already in use will be integrated and if necessary extended. Of course, in a second step the individual corpora are to be classified according to the extended set of metadata.

The metadata should be compatible with existing linguistic metadata standards, especially with IMDI[6] and the the metadata set of OLAC[7]. However in different aspects the new set of metadata will be more specific.

4.2. Integration of terminologies

As already recognized by several researchers the problem of combining existing, real annotation vocabularies with a repository of linguistic categories is a crucial one (Ide et al., 2003). Since we expect standard based solutions to meet the need of sustainability most appropriately, we intend to use and/or to produce a data repository on the base of OWL (McGuinness and van Harmelen, 2004), such as the resource GOLD[8].

We would like to start with an ontology such as GOLD and to successively extend the existing ontology with sub-ontologies[9] for all the annotated phenomena in the projects of the SFBs. Since it has been shown that GOLD is extensible and therefore applicable for diverse kinds of linguistically motivated annotation vocabularies (Goecke et al., 2005), we are quite confident, that GOLD is a good candidate for an appropriate base ontology for linguistic categories. A first study on the integration of the GOLD-Ontology will be presented in (Chiarcos et al., to appear).

Following the LAF proposal, in a second step a mapping from the annotation vocabularies to the ontology will be defined and implemented.

5. Outlook

This paper has focussed on data models, data formats and software tools for sustainable linguistic resources. There are, however, less technical aspects that have an equally relevant impact on sustainability. On the one hand, this concerns the way individual researchers or research projects approach their data handling in the first place - a lot of problems that arise with respect to sustainability of linguistic data could be avoided or at least mitigated if some basic agreement on a set of best practices (e.g. use of open standards and non-proprietary software, or a minimum set of metadata) could be achieved on a broad basis in the research community. Suggestions for such rules of best practice have been made, e.g. (Bird and Simons, 2003), and the project described here intends to elaborate on this work and contribute to its spreading in the research community. On the other hand, insecurities about questions of copyright and of individual rights of persons recorded for linguistic studies often constitute a major obstacle to making linguistic corpora available to a broader public. Here too, the project aims to investigate possible ways of overcoming these obstacles and to formulate rules of best practice. A more comprehensive description of these tasks will be provided in (Chiarcos et al., to appear).

[6] ISLE (International Standard for Language Engineering) Meta Data Initiative, see (Wittenburg et al., 2002)

[7] Open Language Archives Community, see (Bird and Simons, 2004)

[8] General Ontology for Linguistic Description, see (Farrar and Langendoen, 2003)

[9] These specific sub-ontologies are named Community-specific extensions (COPEs) in the GOLD-Terminnology)

6. References

Petra Saskia Bayerl, Harald Lüngen, Daniela Goecke, Andreas Witt, and Daniel Naber. 2003. Methods for the semantic analysis of document markup. In C. Roisin, E. Munson, and C. Vanoirbeek, editors, *Proceedings of the ACM Symposium on Document Engineering (DocEng 2003). pp. 161 - 170)*, pages 161 – 170.

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*.

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79:557 – 582.

Steven Bird and Gary Simons. 2004. Building an open language archives community on the dc foundation. In Diane I. Hillmann and Elaine L. Westbrooks, editors, *Metadata in practice*, pages 203 – 222. American Library Association., Chicago.

Jean Carletta, Jonathan Kilgour, Timothy J. O'Donnell, Stefan Evert, and Holger Voormann. 2003. The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets. In *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML)*.

Christian Chiarcos, Erhard Hinrichs, Timm Lehmberg, Georg Rehm, Thomas Schmidt, and Andreas Witt. to appear. From project data to sustainable archiving of linguistic corpora. In *Paper accepted at the E-MELD workshop 2006*, Ypsilanti.

Stefanie Dipper, Michael Götze, Manfred Stede, and Tillmann Wegst. 2004. ANNIS: A linguistic database for exploring information structure. In Shinichiro Ishihara, Michaela Schmitz, and Anne Schwarz, editors, *Interdisciplinary Studies on Information Structure (ISIS)*, volume 1, pages 245–279. Universitätsverlag Potsdam, Potsdam, Germany.

Stefanie Dipper. 2005. XML-based Stand-off Representation and Exploitation of Multi -Level Linguistic Annota-

tion. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin, Germany.

Scott Farrar and Terry Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International*, 7(3).

Daniela Goecke, Harald Lüngen, Felix Sasaki, Andreas Witt, and Scott Farrar. 2005. Gold and discourse: Domain- and community-specific extensions. In *Proceedings of the E-MELD workshop on Morphosyntactic Annotation and Terminology: Linguistic Ontologies and Data Categories for Language Resources.*, Cambridge, MA.

Michael Götze, Stavros Skopeteas, Thorsten Roloff, and Ruben Stoel. to appear. Towards an exploration infrastructure for a cross-linguistic production data corpus. In *Proceedings of the Sixth International Tbilisi Symposium on Language, Logic and Computation*, Batumi, Georgia.

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

Nancy Ide, Laurent Romary, and Eric de la Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Deborah L. McGuinness and Frank van Harmelen. 2004. OWL Web Ontology Language. Technical report, World Wide Web Consortium.

David McKelvie, Amy Isard, Andreas Mengel, Morten Baun Møller, Michael Grosse, and Marion Klein. 2001. The mate workbench — an annotation tool for xml coded speech corpora. *Speech Communication*.

Thomas Schmidt. 2005a. *Computergestuetzte Transkription - Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln*. Peter Lang.

Thomas Schmidt. 2005b. Time based data models and the text encoding initiative's guidelines for transcription of speech. *Working Papers in Multilingualism, Series B*.

C. M. Sperberg-McQueen and Lou Burnard, editors. 1994. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Chicago, Oxford: Text Encoding Initiative.

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*, Barcelona, Spain.

Andraes Wagner and Bettina Zeisler. 2004. A syntactically annotated corpus of Tibetan. In *Proceedings of LREC 2004*, pages 1141–1144, Lisboa, May.

Andreas Wagner. 2005. Unity in diversity: Integrating differing linguistic data in TUSNELDA. In Stefanie Dipper, Michael Götze, and Manfred Stede, editors, *Heterogeneity in Focus: Creating and Using Linguistic Databases*, volume 2 of *ISIS (Interdisciplinary Studies on Information Structure), Working Papers of the SFB 632*, pages 1–20. Potsdam.

Peter Wittenburg, Wim Peters, and Daan Broeder. 2002. Metadata proposals for corpora and lexica. In *Proceedings of the Third Language Resources and Evaluation Conference (LREC)*, pages 1321 – 1326, Las Palmas.