E-MELD 2006 Workshop on Digital Language Documentation:
Tools and Standards: The State of the Art

# Avoiding Data Graveyards:
# From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources

Thomas Schmidt, Christian Chiarcos, Timm Lehmberg,
Georg Rehm, Andreas Witt, Erhard Hinrichs

**Abstract**

This paper describes a new research initiative addressing the issue of sustainability of linguistic resources. The initiative is a cooperation between three collaborative research centres in Germany – the SFB 441 "Linguistic Data Structures" in Tübingen, the SFB 538 "Multilingualism" in Hamburg, and the SFB 632 "Information Structure" in Potsdam/Berlin. The aim of the project is to develop methods for sustainable archiving of the diverse bodies of linguistic data used at the three sites. In the first half of the paper, the data handling solutions developed so far at the three centres are briefly introduced. This is followed by an assessment of their commonalities and differences and of what these entail for the work of the new joint initiative. The second part then sketches seven areas of open questions with respect to sustainable data handling and gives a more detailed account of two of them – integration of linguistic terminologies and development of best practice guidelines.

# 1. Introduction

In the last two decades, the amount of language data collected and processed for linguistic research purposes has increased dramatically. Most of the time, the data formats and annotation standards as well as the content depend on the research questions a specific project pursues. In conjunction with technological changes, this diversity causes a high degree of heterogeneity that is responsible for the fact that, usually, it is rather difficult to exchange these data collections with other groups or to reuse them in different research contexts when the initial project is completed. This current state of affairs is most unfortunate, since compiling the data requires the investment of extensive (technical and human) resources, often financed by third-party funds.

The joint initiative "Sustainability of Linguistic Data" that we describe in this paper is formed by three research centres, SFB[1] 538 ("Multilingualism"), SFB 441 ("Linguistic Data Structures") and SFB 632 ("Information Structure"), each funded by the Deutsche Forschungsgemeinschaft (DFG). These three centres have collected language data over a period of several years and have processed it according to their specific research questions (see section 2). Taken as a whole, the collection of three data collections contains, for example, a wide range of data types (including written and spoken language), synchronic and diachronic data, hierarchical and timeline-based markup (on several annotation levels), and lexical resources.

The primary goal of our initiative is to convert the data collected by the three collaborative research centres into a comprehensive and sustainable linguistic corpus archive that we aim to be accessible and usable by researchers and applications for at least five decades. In addition, methodologies and rules of best practice for data storage, annotation, and access will be developed. We see our work as a kind of blueprint for comparable initiatives.

The paper is structured as follows: section 2 gives a detailed overview of the general problem (2.1), the preparatory work done so far within the three research centres (2.2 to 2.4), and the specific tools used to access the data collections. Section 3 shows the relevant commonalities and differences between the approaches. Section 4 introduces seven main areas of future work and sketches four of these briefly. The remaining three areas are described in more detail in sections 5 and 6.

# 2. From Project Data to TUSNELDA, EXMARaLDA and PAULA

## 2.1 General Problem

The three research centres involved in this joint initiative bring together researchers sharing a common interest in a linguistic research topic, but also differing in many ways with respect to their individual research backgrounds and aims. A problem visible (and, in some cases,

---

[1] SFB is an acronym of "Sonderforschungsbereich" – a "Collaborative Research Centre".

already acute[2]) from the outset was that these differences would result in highly heterogeneous approaches to linguistic data handling, and that this heterogeneity could potentially hinder cooperation between projects. Such difficulties are well known and have been widely discussed (see, for example, the contributions in Dipper el al. 2005.). In essence, the problem is that researchers often create linguistic data with a specific linguistic theory and a concrete research question in mind. Data formats as well as the tools used to create, edit and analyse corpora are tailored to the specific task at hand, and little attention is paid to the question of how these corpora could be exchanged or reused for other purposes in the future. More often than not, this results in data that is dependent on a single piece of software or on a specific operating system and that becomes difficult to use when this software is no longer supported by its developers. Even where no such fundamental technical obstacles exist, the lack of proper documentation or difficulties in adapting a resource to the requirements of a new research question can greatly hamper data exchange and reuse.

The research centres involved in our joint initiative have addressed these problems right from the start: at each site, a central project is assigned with the task of developing methods for the creation, annotation and analysis of linguistic data that lend themselves more easily to exchange and reuse. The following sections briefly sketch the solutions developed so far.

## 2.2 SFB 441: Linguistic Data Structures

The principal concern of the research centre SFB 441 at Tübingen University are linguistic data structures and their application for the creation of linguistic theories. This general problem is approached from a variety of research perspectives: SFB 441 comprises a total of 12 projects, each of which investigates a specific linguistic phenomenon, either with regard to general methodological issues or concerning a particular language or language family. For example, the research questions range from syntactic structures in German and English, local and temporal deictic expressions in Bosnian, Croatian, Serbian, Portuguese and Spanish, to semantic roles, case relations, and cross-clausal references in Tibetan.

### 2.2.1. Corpora and Tools

Many SFB 441 projects create digital collections of linguistic data as the empirical bases for their research and prepare them to fit their particular needs. Usually these collections are text corpora. In addition, a couple of projects deal with data (e.g., lexical information) that are more adequately represented by an Entity-Relationship based data model, implemented in relational databases. All SFB 441 data collections are compiled in a single repository called TUSNELDA. The corpora are integrated into an XML-based environment that ensures common methods for encoding, storing, and retrieving data. This integration is particularly challenging due to the heterogeneity of the individual corpora: they differ with regard to properties such as language (e.g., German, Russian, Portuguese, Tibetan), text type (e.g., newspaper texts, diachronic texts, dialogues), informational categories covered by the annotation (e.g., layout, text structure, syntax), and underlying linguistic theories (see Wagner, 2005, for an overview). The size of the individual corpora ranges from 10,000 (Spanish/Portuguese spoken dialogues) to ca. 200 million words (automatically chunk-parsed German newspaper texts). Several tools are in use to capture and process the data: for example, treebanks are built using Annotate, the XML editor CLaRK is used for the annotation of Tibetan texts (e.g., text structure, and morphological features), and prototypes of Web-accessible querying interfaces were implemented using Perl scripts as well as the native XML database Tamino.

---

[2] Large amounts of the data used and analysed in several projects have been collected in previous projects, i.e. long before the respective research centre was founded.

## 2.2.2. Data Format

In spite of the diversity of the corpora contained in the TUSNELDA repository, they all have in common the same generic data model: hierarchical structures. It is most appropriate to encode the phenomena researched in the SFB 441 projects by means of nested hierarchies, occasionally augmented by secondary relations between arbitrary nodes. This key property distinguishes the TUSNELDA collection fundamentally from speech corpora annotated with regard to timeline-based markup or frommultimodal corpora. Such corpora usually encode the exact temporal correspondence between events on parallel layers (e.g., the coincidence of events in speech and accompanying gestures, or the overlap of utterances), whereas hierarchical aspects are of secondary interest only. In TUSNELDA, however, hierarchical information (e.g., textual or syntactic structures) is prevalent. As a consequence, the TUSNELDA annotation scheme encodes information according to the paradigm of embedded (rather than standoff) annotation, directly resulting in hierarchical structures (the trees created by nested XML elements). The decision to employ the hierarchical paradigm is primarily based on the fact that this procedure makes it possible to utilise off-the-shelf XML-enabled tools (such as XML editors, filters, converters, XML databases, and query engines). In addition, whenever a tool that has already been in active use in one of the projects was unable to export an XML-format, Perl scripts and XSLT stylesheets have been used to transform the legacy data into TUSNELDA's XML-based format.

The structures encoded in the TUSNELDA corpora do not overlap and can be integrated into a single hierarchy. For example, syntactic structures constitute sub-sentential hierarchies, whereas text structures define super-sentential hierarchies. Structures of this kind can be captured within a single XML instance. Overlapping structures are very uncommon and, therefore, they are not of primary importance. These units concern the annotated texts' layout structure such as page boundaries. Boundaries of this kind are marked by milestone elements (e.g., <pb/> for a page break) that do not violate the well-formedness of the XML document (see Wagner and Zeisler, 2004, for details).

## 2.3 SFB 538: Multilingualism

The SFB 538 "Mehrsprachigkeit" (Multilingualism) took up its work in 1999. It currently consists of 14 projects doing research on various aspects of multilingualism, the most important of which are bilingual first language acquisition, multilingual communication and historical multilingualism. Researchers come from a variety of backgrounds with generative grammar and functional approaches (functional pragmatic discourse analysis, systemic functional linguistics) being the dominant paradigms. Languages studied include Germanic and Romance languages (each also in their historic stages), Turkish, Basque and sign language.

## 2.3.1. Corpora and Tools

All projects work with empirical data. For the greater part, this means corpora of transcribed spoken interaction, most importantly child language acquisition data and other spontaneous conversational data. Corpora of written language are mainly used in projects with a diachronic perspective on multilingualism.

When the research centre started its work, several researchers had already collected large amounts of linguistic data that had to be integrated into the new collections. This extensive set of legacy data was created with a diverse set of transcription and annotation tools:

- syncWriter – a Macintosh tool for creating data in musical score ("Partitur") notation
- HIAT-DOS – a similar tool (for MS Windows)
- Wordbase – a 4th-Dimension database software application (for Macintosh computers)

- LAPSUS – a dBase III database application (for MS Windows)

In their original form, these data collections were entirely incompatible with one another. Even though syncWriter and HIAT-DOS data on the one hand, and Wordbase and LAPSUS data on the other are conceptually very similar, their dependence on a specific software (and thereby on the operating system on which this software runs) made even basic processes such as viewing the data on a different machine an impossible task. Moreover, since the software tools in question were no longer supported by their developers, it was anticipated that the corpora will, in the medium term, become unusable – even for their original creators. As a consequence, a central project was funded with the task of developing a solution that would make the collections more sustainable and more readily exchangeable. The EXMARaLDA system, presented in the next section, was developed in this project.

Due to the amount of manual work involved in the process, conversion of legacy data is still ongoing. Nevertheless, the majority of the research centre's spoken language data are now available in EXMARaLDA XML. Corpora for which the conversion work has been almost completed include: a corpus of conversational data from Turkish/German bilingual children; a corpus of Scandinavian semi-communication (mostly radio broadcasts involving a Danish and a Swedish native speaker); a corpus of interpreted (German/Portuguese and German/Turkish) doctor patient communication – all transcribed according to discourse analytical principles; a phonetically transcribed corpus of acquisition data from Spanish/German bilingual children. New corpora, i.e., corpora created in the EXMARaLDA framework include a corpus of simultaneous and consecutive interpretation between German and Portuguese; a phonetically transcribed corpus of Catalan; and a corpus of semi-structured interviews with bilingual speakers of Faroese.

All in all, the research centre's data will contain more than 1,000 hours of transcribed speech in different languages and from different domains.[3] Added to this are a number of written language corpora, most of which are also in a (TEI compliant) XML format.

## 2.3.2. Data Format[4]

EXMARaLDA defines a data model for the representation of spoken interaction with several participants and in different modalities. The data model is based on the annotation graph approach (Bird/Liberman 2001), i.e., it departs from the assumption that the most important commonality between different transcription and annotation systems is the fact that all entities in the data set can be anchored to a timeline. EXMARaLDA defines a basic version of the data model which is largely similar to other data models used with software for multimodal annotation (e.g., Praat, TASX, ELAN, ANVIL). This has proven an appropriate basis for the initial transcription process and simpler data visualisation and query tasks. An extended data model that can be calculated automatically from the basic version by exploiting the regularities defined in transcription conventions caters for a more complex annotation and analysis.

Data conforming to this model is physically stored in XML files whose structure is specified by document type definitions (DTDs). Conversion filters have been developed for legacy data (see above). Due to a lack of documentation and several inconsistencies in these older corpora, however, a complete conversion cannot be accomplished automatically, but requires a substantial amount of manual post-editing.

New data is now usually created as EXMARaLDA data with the help of the EXMARaLDA Partitur-Editor, a tier-based tool presenting the transcription to the user as a musical score and supporting the creation of links between the transcription and the underlying digitized audio

---

[3] A more exact estimate is difficult at this point in time because many corpora are still growing in size.
[4] A more detailed account of the EXMARaLDA data model is given in Schmidt (2005a,b). The EXMARaLDA tools are described in detail in Schmidt/Wörner (2005) and in various materials available from the project website (http://www.rrz.uni-hamburg.de/exmaralda)

or video recording. Alternatively, compatible tools like ELAN, Praat, or the TASX annotator can be used to create EXMARaLDA data. The EXMARaLDA corpus manager is a tool for bundling several transcriptions into corpora and for managing and querying corpus metadata. ZECKE, the prototype of a tool for querying EXMARaLDA corpora, is currently evaluated.

## 2.4 SFB 632: Information Structure

SFB 632 "Information Structure" is a collaborative research centre at the Humboldt University of Berlin and the University of Potsdam. Established in 2003, it currently consists of 14 projects from several fields of linguistics.
The variety of languages examined is immense and covers a broad range of typologically different languages and language stages (e.g., several Indo-European languages, Hungarian, Chadic, Georgian, Japanese, etc.). In the research centre, integrative models of information structure based on empirical linguistic data are developed. Thus, most projects make use of data collections such as linguistic corpora (e.g., the Potsdam Commentary Corpus, Stede 2004), collections of elicited speech (Questionnaire for Information Structure, Féry et al. 2006), or experimental data.

### 2.4.1. Corpora and Tools

As the projects focus on the interaction between information structure and other structural levels of language (e.g., phonology, syntax, discourse structure), the corpora are characterised by multiple levels of annotation. For these corpora, the SFB632 annotation standard[5] is applied which includes guidelines for the annotation of morphology, syntax, semantics, discourse structure, and information structure. The standard was developed by interdisciplinary working groups that involved researchers from theoretical linguistics, psycholinguistics, phonology, historical linguistics, computational linguistics and comparative linguistics. Accordingly, the standard and the guidelines are designed under the assumptions of language-independence and generality. Furthermore, PAULA, the "Potsdam exchange format for linguistic annotation" (Potsdamer Austauschformat für Linguistische Annotationen) has been developed. PAULA is a generic XML format capable of representing the full variety of annotations in the research centre. Import scripts for several source formats exist, e.g., from EXMARaLDA, MMAX, MMAX2 (Müller and Strube 2001), RS3 (RSTTool, O'Donnell 1997), and URML (Reitter and Stede 2003).
PAULA is the primary input format for ANNIS ("annotation of information structure"), the SFB 632 database application (Dipper et al. 2004, Dipper and Götze 2005). ANNIS is a web application that can be used to view and to search corpora. The tool is still under development. A web demo will be available soon to registered users at http://www.sfb632.uni-potsdam.de/annis. It comprises samples of the following corpora:

- samples of 13 typologically different languages from the "Questionnaire for Information Structure" (Féry et al. 2006),
- the Potsdam Commentary Corpus of German newspaper commentaries (Stede 2004), and
- parts of the Old High German corpus described in (Hinterhölzl and Petrova 2005).

In addition to its function as the primary input format for ANNIS, PAULA is applied for automatic text summarisation and statistical analyses. It is intended to integrate all annotations assembled by the SFB projects, thus, we chose a distributed multi-level, or "standoff" architecture (see below). This is necessary as the SFB 632 corpora are typically annotated on multiple levels, often with overlapping hierarchies or segmentation. As an

---

[5] The SFB632 annotation standard is currently being prepared for publication.

example, the Potsdam Commentary Corpus (Stede 2004), a collection of 175 German newspaper commentaries, is annotated for parts of speech (STTS, Skuts et al. 1998), syntax (TIGER, Brants et al. 2002), discourse or rhetorical structure (URML, Reiter and Stede 2003), anaphoric co-reference (PoCoS, Chiarcos and Krasavina 2005), and partially, for detailed morphological information, information structure (SFB632 annotation standard), and discourse connectives.

### 2.4.2. Data Format

Since very richly annotated corpora are considered as necessary for research on Information Structure, it becomes inevitable that segments on different structural levels (e.g., markables from coreference annotation and constituent structures from morphosyntax) cannot be harmonised in a common representation by means of a single XML hierarchy. As a possible solution, PAULA is a generic XML-based format, which is inspired by the annotation methodology developed in the Linguistic Annotation Framework (Ide et al. 2003) and is designed as standoff-architecture.
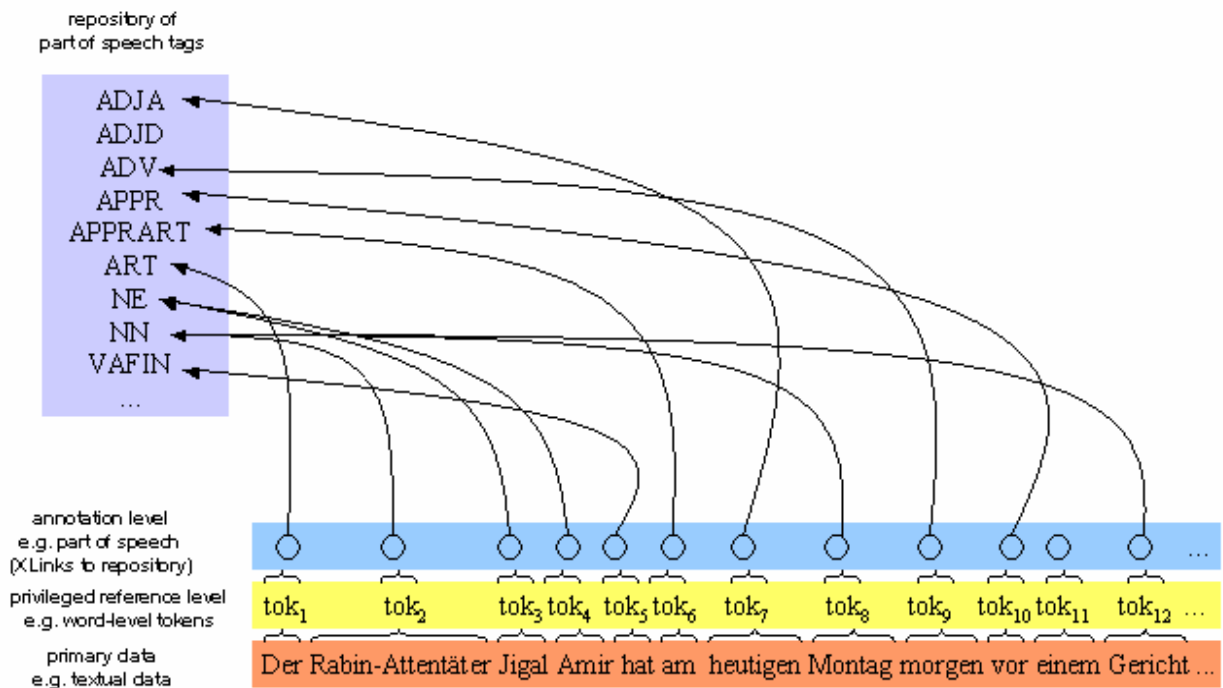


**Figure 1: Standoff annotation in PAULA. Annotation of part of speech tags with reference to a central data repository.**

"Standoff architecture" describes an approach that stores data (e.g., a text file or an already annotated text) separate from (additional) annotations. Hence, with standoff annotation it is possible to distribute text data and annotations over separate files,[6] so that the file containing the source text can be left untouched. This approach allows for the annotation of text that cannot be modified (e.g., because the text is stored on a read-only medium). Moreover, whereas XML does not easily account for overlapping segments and conflicting hierarchies, they can be marked up in a natural way in standoff annotation by distributing annotations over multiple layers, possibly interlinked with one another. Therefore, not only might the primary data be separate from the annotation, but individual annotations can be kept apart from one another in independent files as well.

Currently, PAULA is focused on textual information, i.e., the central reference level (or "primary data") is a privileged token level that is either the textual primary data, or the

---

[6] But it is also possible to include all the annotations in one XML-document.

transcription of video or audio data. It is possible, however, to refer to structures beyond this level instead of directly relating annotations to the reference level alone, thus allowing for stratal and hierarchical annotation. Possible extensions of the format might permit the integration of an additional, more primitive level (e.g., timelines or position in an audio file) that the primary data refers to. Furthermore, PAULA employs a strict separation of annotation and annotation scheme. Features are not embedded into the annotation, but they refer to a central repository.

As the PAULA format does not depend on XML hierarchies it allows for the representation of non-nesting hierarchical structures and arbitrarily complex relationships, including time-line information (Dipper et al. 2006), anaphoric or bridging relations (PoCoS, Chiarcos and Krasavina 2005).

PAULA is still work in progress. Currently, an inline format is developed that enables more efficient processing, as the extensive use of XLink references can be considered very problematic from a performance point of view. Similarly, the format is certainly not suitable for human inspection and debugging. The inline representation is generated by parameterised scripts using milestones as a fallback for cases of overlapping hierarchies. For a similar solution, see Witt (2004).

## 3. Linguistic Data Processing at the Three Sites: A Comparison

The manner in which the problems of data exchange and reuse have been addressed at the three sites share some obvious common characteristics. The general approach of TUSNELDA, EXMARaLDA and PAULA is to convert data from a variety of project-specific formats to a more universal XML format based on some abstract model of linguistic data. This data model, in turn, acts as the core architectural component of a framework in which – ideally – researchers can retain a project-specific perspective towards their "own" data while, at the same time, they are able to reuse other projects' resources. Besides facilitating data exchange in the present, the use of standardised technologies such as XML and Unicode as well as the fact that the three generalised data models are more thoroughly documented than their project-specific predecessors also enhances the long-term usability of resources.

However, the three approaches also differ in some important aspects. This is perhaps most obvious in the specifics of the data models and formats. Firstly, different text-technological frameworks have been used as a starting point in the development of the three formats. TUSNELDA is based on the work of the XCES initiative (EAGLES 2000) which, in turn, was inspired by the guidelines of the Text Encoding Initiative (Sperberg-McQueen/Burnard 1994). EXMARaLDA uses the approach suggested by Bird/Liberman (2001) in their annotation graph framework. Finally, PAULA is most directly related to the standoff annotation methodology developed in the Linguistic Annotation Framework (Ide et. al 2003). These origins entail two important differences in the solutions developed:

- Whereas PAULA is committed to a standoff-annotation approach, i.e., a strict separation of annotation levels and their encoding in separate documents, TUSNELDA favours an integrated representation, i.e., single documents in which a number of different annotation levels can be included.[7] EXMARaLDA is standoff insofar as it keeps separate annotation levels on separate tiers, but otherwise also favours an integrated, single-document approach.
- TUSNELDA is a hierarchy-based data model. It sees hierarchies as the primary relation in linguistic data sets and treats other (e.g., temporal) relationships as secondary. EXMARaLDA takes the opposite approach by privileging temporal over

---

[7] Following a level/layer distinction made by Bayerl et al. (2003), the notion of "annotation level" refers to an abstract level of analysis (e.g., morphology, rhetorical structure, document structure, syntax). The terms "layer" or "tier" refer to a concrete realisation by means of, for example, an XML file.

hierarchical relations[8]. PAULA, finally, follows a hybrid approach in which neither hierarchical nor time-based relations are privileged.

Of course, these differences are primarily motivated by the diverse needs of researchers at the three sites and, consequently, by dissimilar priorities of the projects which have been responsible for devising a common data format. In the study of multilingual language acquisition and communication, spoken data (often with many participants and sometimes in more than one modality) are a much more common source of empirical studies than written data[9]. The most pressing need in Hamburg has therefore always been the development of appropriate transcription tools, while the handling of written data is given lower priority. Consequently, EXMARaLDA is relatively well-suited for earlier steps in the creation of spoken language corpora, whereas written corpus development, extensive data annotation and complex analyses of several annotation levels are not (yet) equally well supported. In Tübingen and Potsdam/Berlin, on the other hand, no equally strong bias towards spoken data exists, so that both TUSNELDA and PAULA cater for written and spoken data. Tool development at these sites focuses more on the annotation of existing resources and on query methods. In a way, the solutions developed thus address complementary needs: data creation in Hamburg, data annotation and query in Tübingen and Potsdam/Berlin. Work package 1 of the joint initiative (see section 4) will be concerned with finding ways of reconciling these complimentary solutions.

For similar reasons, data resources in Tübingen and Potsdam/Berlin are already partly integrated in central repositories (the TUSNELDA and ANNIS databases), whereas the Hamburg data, although now represented in a common format, still lack this kind of integration. The work packages 2, 3 and 4 are mainly concerned with issues in this area.

Finally, another important difference of the three approaches is the degree to which the sites have attempted a standardisation of the linguistic data categories themselves. In Hamburg it was felt that the theoretical backgrounds and research aims of the different projects were too diverse to attempt a standardisation of linguistic categories across all projects.[10] EXMARaLDA was therefore explicitly designed as an "ontologically empty" framework (a framework that generalises over theory-neutral characteristics of different linguistic data sets, but otherwise makes no claim to reconcile different theoretical approaches).[11] In both Tübingen and Potsdam/Berlin, on the other hand, some effort has been made to agree not only on a common abstract structural backbone for all data sets, but also on some common way to treat and encode a set of specific linguistic phenomena. Site-wide annotation guidelines (Wagner/Kallmeyer, 2001, and Chiarcos/Krasavina, 2005) are the result of these efforts. These guidelines predetermine to a great part the work on terminology integration to be done in work package 5.

---

[8] See Schmidt (2005a,b) for a more detailed discussion of the relationship between hierarchy- and time-based data models.

[9] However, researchers in Hamburg also work with written language corpora, most importantly in those projects interested in diachronic aspects of multilingualism. Since EXMARaLDA is not meant to cater for these resources, the general strategy for these projects has been to rely on the TEI guidelines as much as possible. One annotation tool – the Z2-Tagger (see http://www.exmaralda.org) – has been developed for this kind of data.

[10] Consider, for instance, the notion of an "utterance". For a researcher working on phonetic aspects of L1 acquisition in a 15 month old child, this category is motivated by very different considerations than for a discourse analyst investigating turn taking in adult expert communication. It would be possible, of course, to identify what little commonality the two notions have and integrate that "knowledge" into the annotation framework, but the benefits of that exercise may be doubted.

[11] A theory-specific annotation guideline (Rehbein et al. 2004), however, was developed for the projects that had previously used slightly different variants of the HIAT transcription system.

**Table 1: Key properties of the three annotation formats used in the Research Centres in Tübingen, Hamburg, and Potsdam/Berlin**

| | TUSNELDA (SFB 441, Tübingen) | EXMARaLDA (SFB 538, Hamburg) | PAULA (SFB 632, Potsdam/Berlin) |
|---|---|---|---|
| Languages (selection) | German, Russian, Portuguese, Tibetan, | Japanese, German, Turkish, Portuguese, English, Swedish, Danish, Norwegian, Italian, French, Basque, Greek, Faroese, Sign Language | German, Latin, several Gur- and Kwa-languages, several Chadic languages |
| Key topic of the Centre: | Linguistic data structures | Multilingualism | Information structure |
| Representation format: | XML | XML | XML |
| Text-technological foundation: | XCES, TEI | annotation graphs (Bird, Liberman, 2001) | stand-off annotation (Linguistic Annotation Framework) |
| Annotation levels: | integrated representation in a single file | annotation levels are kept on separate tiers but in a single file | separation of annotation levels, encoding in multiple files |
| Primary relation in linguistic data sets: | hierarchies (i.e., trees) | temporal data | hybrid |
| Data sources: | primarily written data | primarily spoken data | primarily written data and transcriptions |
| Tool development: | annotation of existing resources, querying | transcription, creation of spoken language corpora | annotation of existing resources, querying |
| Annotation guidelines: | yes | No | yes |
| Central data repository: | yes | no | yes |

# 4. From TUSNELDA, EXMARaLDA, and PAULA to Sustainable Archives

Based on an assessment of the current state of linguistic data processing as described in section 3, our joint initiative has identified seven areas of open questions towards genuinely sustainable linguistic archives. Four of these will only be briefly sketched here – Dipper et al. (2006) contains a more detailed discussion of these points. The remaining three areas will be described in somewhat more detail in sections 5 and 6.

## 4.1. Development of Data Formats

Just as TUSNELDA, EXMARaLDA and PAULA set out to generalise over a number of project-specific data models and formats, our joint initiative has as one of its primary goals the development of a data model which generalises over TUSNELDA, EXMARaLDA and PAULA. Although XML and Unicode can in all three cases be regarded as some form of base level annotation, the abovementioned task remains non-trivial as it requires a reconciliation of conceptually different text-technological paradigms. We are currently exploring to what

extent the NITE Object Model (NOM, Carletta et al. 2003) can be used as a starting point for this task.

## 4.2. Development of Methods and Tools for Data Distribution and Data Access

Once linguistic resources are available in a form that makes them suitable for reuse in other research contexts, archiving and disseminating methods have to be developed. Although an XML-based format such as the one we are aiming at seems to be a promising candidate for achieving long-term usability for digital data, there is no guarantee that such data will remain accessible in the very long term (say, over the next five decades). We therefore plan not to rely exclusively on digital archiving methods, but also to produce a human-readable hardcopy of the data that can be archived in libraries using their conventional, well-established methods for printed material. Since all the data will be available in an XML format, XSLT and XSL-FO are the obvious candidates for generating printed versions of the corpora. Nevertheless, the option of generating additional formats, especially the Open Document Format and Office XML, will also be considered.

Short and medium term data dissemination, on the other hand, has to focus on methods that allow researchers quickly to discover an existing resource, to assess its relevance for their research purposes and to obtain it in a form that is suited to the working processes used in the project. For the first two steps, web interfaces offering access to (possibly distributed) corpus documentation and allowing a query of corpus metadata are the most promising approach. Web-based solutions can have advantages and disadvantages. One advantage is that they usually require no specific software tools and no local storage capacity (this is especially relevant when corpora are not exclusively text-based, but also contain storage-intensive audio or video material). Other reasons, however, may make it preferable to have offline access to the data also – for example, when only slow dial-up Internet access is available or when a researcher has to transform a complete resource in order to work with it using locally installed tools. We therefore plan to develop both web-based data distribution methods and digital versions of corpora to be distributed on offline media such as DVDs.

## 4.3. Query Interfaces

Online and offline data distribution methods will have to provide means of querying both metadata and the linguistic data itself. In terms of sustainability, it is crucial that such query mechanisms be intuitively usable by a wide range of researchers, as the usability of interfaces has a decisive impact on whether and how a resource can be reused. Moreover, it is desirable that similar query tasks be approachable in a similar manner across heterogeneous data sets. For instance, the task of string pattern matching occurs in almost every corpus analysis, and it is in the interest of sustainability that users do not have to learn a new query syntax whenever they use a new corpus. For the same reasons, it is important that query mechanisms, like the data itself, are based on technologies that are widely accepted and, ideally, managed as open standards, so that there is at least some kind of guarantee that they will still be usable in a decade or two.

As indicated in section 2, the solutions developed at the three sites already provide a number of query mechanisms for the respective data models. In our joint initiative, we will build upon this work, adapting and extending these solutions to the data format described above. Because issues of efficiency (e.g., short response time) and expressive power (i.e., the possibility of formulating structurally complex queries) can often run counter to the goal of sustainability, we will consciously background these issues wherever they conflict with our primary goal. For reasons cited above, standardised technologies such as XSLT and XQuery will be used as the technological basis for new query tools.

### 4.4. Data Integration

As long as a data resource is only used within the project that created it, often only minimal attention is paid to systematically and explicitly recording metadata (i.e., information about the general composition of the corpus, details about speakers of transcribed interactions or authors of written texts, etc.), because the researchers have been involved directly in the creation of the corpus and therefore have such information "in their heads". However, when the resource is to be made available to other researchers, it becomes crucial for such information to be represented in a systematised, digital form. Firstly, so that the value and relevance of a resource can be assessed without studying every detail. Secondly, because certain types of metadata (e.g., a speaker's age or gender) may be immediately relevant for certain corpus analyses.

Our joint initiative therefore plans to compile a comprehensive set of metadata. Since this set must adequately describe all the corpora of all the research centres, existing metadata will have to be integrated and extended where necessary. In a second step, such an extended set of metadata will be used to derive a classification of the different corpora. As with the issues mentioned in previous sections, compatibility with existing standards is a key requirement for sustainability in the documentation of metadata. The most important standards in this area are IMDI[12] and the metadata set of OLAC[13].

## 5. Integration of Linguistic Terminology

### 5.1. Diversity of Data

One of our primary aims is to provide the means to ensure the long-term availability of the data collections. Along with technical aspects, as discussed by Dipper et al. (2006), this goal involves creating a thorough documentation for the corpora in order to provide easy access for non-specialised users. This includes metadata about the corpora themselves, such as type of data, formats, standards and levels of annotation. Furthermore, the terminology relevant for the annotations has to take into account sustainability considerations. Annotation details such as definitions of tags, tag names, labels of syntactic structures, etc., can be highly idiosyncratic or restricted to the conventions of a specific community.

As an example, consider part of speech (POS) annotation. For historical reasons, it is common practice for English to mark adjectives with tags beginning with J,[14] whereas most modern tagsets use abbreviations derived from grammatical terms from Latin or English, (e.g. Menota: A, STTS: ADJA, ADJD), but occasionally from other languages as well (e.g. PRIL for a Russian corpus [Roland Meyer, p.c.], from Russian *imya prilagatel'noye*.). Similar idiosyncrasies or language-specific design elements can be found in the definitions of tag names and annotated structures. As a consequence, researchers in typological or comparative studies might be hindered by community-specific term definitions. Further, for better studied languages such as English and German, several incompatible tagsets have been applied. Therefore the effort required for large scale studies involving different data sources increases if corpora with different tagsets are involved.

For researchers unfamiliar with the specific usage and origins of terms that have been applied in the creation of a data source such as a corpus, the variety of abbreviations, terms, tags and possibly conflicting definitions can be confusing and time-consuming. In a worst case

---

[12] See http://www.mpi.nl/IMDI/
[13] See http://www.language-archives.org/
[14] Brown: JJ, JJ$, JJR, etc.; London-Lund: JA, JB, JE, etc.; LOB: JJ, JJB, JJR, etc.; Penn: JJ, JJR, JJS; SUSANNE: JA, JB, JB0, etc.

scenario, the effort necessary for a closer examination of the data will prevent later generations of researchers from working with a data collection. The problem becomes even more apparent for very large collections of heterogeneous corpora. That is why it is an urgent task for the unified treatment of such collections to identify and to document commonalities as well as differences in the terminology used: the integration of information on the linguistic terminology can be seen as a core aspect of sustainable maintenance of linguistic data.

Following, we illustrate this problem with the help of the tagsets used in Tübingen, Hamburg and Potsdam/Berlin for POS annotation.

Our research centres create and use POS-annotated corpora for 29 languages or language stages from practically all parts of the world.[15] These corpora are annotated according to nine tagsets or tagset variants: the SFB632 annotation standard (applied for 13 languages), SUSANNE (for English, Sampson 1996), MENOTA (for Old Norse, Old Danish and Old Swedish),[16] a Tibetan tagset (Wagner and Zeisler 2004), a reduced tagset applied to 5 languages for language acquisition studies (further referred to as tagset SFB538/E2, Jasmine Bennöhr, p.c.), a tagset for Russian in three variants (Roland Meyer, p.c., Michael Betsch, p.c.),[17] and three versions of STTS (for German, Skut et al. 1998).

With this amount of data, several problems can be identified that hinder the direct access to data by using these tagsets:

- Partly, tag names are cryptic or just arbitrary, cf. "DD *yon, yonder* as determiner, … DDf *enough ...,* DDi *some,* …, DDo *a lot*, DDy *any*" (Sampson 1996, p. 106), etc.
- Even if tags can be interpreted as abbreviations, idiosyncratic variants in different dialects of the same tagset prevent a direct application of the "canonical" tag name, cf. the tags for pronominal (or "prepositional") adverbs in German in different dialects of STTS: PROP (Tübingen variant), PAV ("canonical" Stuttgart variant), PROAV (TIGER variant).
- Though tag names are usually based on Latin or English grammatical terminology, language-specific tag names are used occasionally.

Besides such "surface" problems, different community- or project-specific definitions of tags and terms can affect their applicability.

- For example, in STTS and SUSANNE, "numbers" are defined differently. In SUSANNE, a "semantic" point of view has been taken (all numerical expressions have tag names that begin with M), whereas STTS introduced an additional syntactic constraint, i.e., only ordinal numbers are tagged as numbers, but cardinal numbers are adjectives.
- A similar problem exists with auxiliary verbs in German. In German, *haben* and *sein* can be auxiliary verbs ("to have"/"to be") or lexical verbs ("to own"/"to exist"). Accordingly, the intuitive interpretation of the STTS tag VAFIN is restricted to the non-lexical reading. Nevertheless, both uses are tagged as VAFIN in both contexts. In SUSANNE, similar surface-based definitions occur.
- Occasionally, tag definitions are extremely complex (e.g. "NNLc L+M+C noun, e.g. *barracks links works*" Sampson 1996, p. 112).
- Finally, tag definitions can be missing completely. See the EAGLES tagsets (Leech and Wilson 1996) for typical examples.

---

[15] The languages and language stages for which POS annotations exist in the SFBs are (the Ethnologue code given in parentheses): Balti (BFT), Basque (EUS), Bole (BOL), Old Danish (n/a), Dutch (NLD), English (ENG), French (FRN), Georgian (KAT), German (GER), Old High German (n/a), Greek (ELL), Guruntum (GRD), Haussa (HAU), Hungarian (HUN), Indonesian (IND), Japanese (JPN), Javanese (JAV), Konkani (KNN), Ladakh (LBJ), Maung (MPH), Niue (NIU), Portuguese (POR), Prinmi [Pumi] (PMI), Russian (RUS), Spanish (SPA), Old Swedish (n/a), Tangale (TAN), Teribe (TFR), and Old/Classical Tibetan (n/a).
[16] www.hit.uib.no/menota/guidelines/, May 4th, 2006.
[17] The Russian tagset is based on the Czech version of MULTEXT-East, cf. http://nl.ijs.si/ME/, May 6th, 2006.

The tagsets are of differing granularity. For example, SUSANNE provides 27 different tags for proper nouns and 47 tags for common nouns in English. In the Penn Tagset[18] a total of four tags exists for proper nouns and common nouns. Similarly, STTS makes a distinction between common nouns and proper nouns for German, which is conflated by the SFB538/E2 tagset. Accordingly, the definition of an elementary and seemingly intuitive term such as "noun" depends on specific knowledge of the specific tagset applied.

The minimal solution to overcome these problems is to provide a consistent terminology and to refer to this terminological backbone in the definition of annotation structures.

## 5.2 The Standardisation Approach

One appealing solution to the problem is the "standardisation approach" as employed by the EAGLES recommendations (Wilson and Leech 1996): The Expert Advisory Group on Language Engineering Standards (EAGLES) was an initiative of the European Commission which aimed to develop

- (i)     standards for large-scale language resources (such as linguistic corpora),
- (ii)    specifications for mark-up languages and software tools and
- (iii)   means for maintenance, assessing and evaluation of resources, tools and products.

With regard to the first aspect, standards for POS tag sets have been formulated which are further referred to as the "EAGLES meta scheme" (Leech and Wilson 1996). These standards are intended to increase tagging accuracy and comparability of automatic taggers and tagsets for most European languages. In a bottom-up approach, existing tagsets for several European languages have been considered, and commonly used terms and categories have been identified. As a result, 13 obligatory categories (noun, verb, adjective, adverb, numeral, pronoun/determiner, article, adposition, conjunction, interjection, unique ["particle"], punctuation, residual) were identified. For each category, a list of features has been assembled that a standard-conformant tagset should respect. Accordingly, the "EAGLES meta tagset" is constituted as the set of "meta tags", i.e. reasonable combinations of categories (main tags) and features.

In general, the standardisation approach can be characterised by three principles:

- *surface-oriented*: The meta tagset is based on the meta scheme, which provides a standardised list of terms, but not of term definitions.
- *direct mapping*: Standard-conformance means that a homomorphism between a tagset and a subset of meta-scheme tags exists:
    1. It *must* provide all obligatory categories of the meta scheme.
    2. It *should* consider all recommended features of the meta scheme.
    3. It *may* include optional features of the meta scheme.
    4. It *may not* include terms beyond this.
- *one-to-many mapping*: Tags in standard-conformant tagsets can correspond to multiple alternative tags in the meta scheme, but not vice versa.

Note that the EAGLES meta scheme is intended to standardise tagsets: it provides a only *technical solution* to the problem of diversity of tagsets. The terminology used is just the sum of terms occurring in the integrated tag sets, but as these are from selected European languages only, it does by no means provide a universal terminological backbone.

From this constellation several problems arise: First, language-specific conceptualizations have to be integrated into the meta-scheme (otherwise, condition 4 of the direct mapping

---

[18] http://www.computing.dcu.ie/~acahill/tagset.html, May 6th, 2006

principle cannot be fulfilled). As an example, the feature "-ing-Form" had to be integrated due to the merging of gerund and participle which is specific to English. As a consequence, the complexity of every standard-conformant scheme is projected onto the meta-scheme.

Secondly, the outcome of the bottom-up process was not a full terminological resource, but only a list of terms. As long as no definitions are included in the description of the standard, community-specific usage of terms can lead to contradictory interpretations of the corresponding tags. Thus, different phenomena are referred to by the same tags. This certainly contradicts any effort of standardisation. Hughes et al. (1995) noted:

> "Our conclusion for the EAGLES Initiative is that the morphosyntactic category proposals must be followed up with detailed definitions, preferably including computable criteria. ... Otherwise the 'standards' will be interpreted differently (and incompatibly) in different tagged corpora. We had hoped that the EAGLES tagset might constitute an 'interlingua' for translating between existing tagsets. However, we have already had to conclude that our task of automatic tagsetmapping extraction can never achieve perfect accuracy, as both source and target training data are noisy; using a fuzzy-edged tagset as an interlingua could only worsen matters."

Finally, this solution is not scalable as it cannot be applied directly to non-European languages (see Khoja et al. 2001 for Arabic). If these languages, however, are to be integrated into the standard, the meta-scheme is forced to keep on growing as more and more language-specific terms, categories or classifications need to be integrated as recommended or optional features.[19] Likewise, the number of obligatory features is forced to decline in this scenario. If the EAGLES meta-scheme was intended for the subsumption of Inuit languages, neither the obligatory categories "article" nor "adposition" would be applicable, and the existence of the category "adjective" could be questioned (Nowak 2005). Furthermore, even elementary categories such as "verb" and "noun" have been questioned for languages such as Tongan (Broschart 1997).

**5.3 Towards a Well-Defined Terminological Backbone**

Two conditions have to be fulfilled to overcome the shortcomings of the standardisation approach.

1. Instead of a simple list of undefined terms, tagsets have to refer to a well-defined terminological backbone.
2. To prevent the projection of complexity from the existing tagsets onto this terminological backbone, the direct mapping requirement between standard-conformant tagsets and a subset of the meta-tagset has to be weakened. As a consequence, universality and scalability of the backbone are enhanced, whereas the complexity of the mapping is higher than in the standardisation approach.

The first condition guarantees terminological consistency, the second condition provides means for the scalability of the complexity reduction of the terminological backbone. Another condition is the existence of an explicit hierarchical structure. Though EAGLES allowed for many-to-one mappings to represent underspecification or absence of a specific feature in a tagset, this could be more intuitively modelled by a hierarchical structure. Especially in the case of "Pronoun/Determiner", with the sub-feature "Category" which allows for a

---

[19] As an example, the MULTEXT-East scheme is an extension of the EAGLES meta scheme to Eastern European languages. Solely due to the integration of Estonian and Hungarian it has to consider 21 additional feature values for case, cf. Ejavec (2004).

specification of "Pronoun" or "Determiner", a hierarchical representation seems to be more natural.[20]

As ontologies provide means for well-defined, structured terminological resources, it seems that these conditions can be most easily fulfilled by the application of an ontology similar to the GOLD approach (Farrar/Langendoen 2003). In contrast to the EAGLES initiative, which was dedicated to European languages exclusively, in the E-MELD project GOLD aspects of universality and scalability were emphasized from the beginning. Instead of providing a generalisation of tagsets for a fixed range of languages, it aimed to cover the full typological variety as far as possible. Finally, it took a different starting point due to its orientation towards the documentation of endangered languages.

As opposed to this, our joint initiative aims to achieve a unified representation and access to existing resources, which – in their quantitative majority – deal with European languages. Accordingly, we suggest to develop an ontology based on established meta-schemes such as EAGLES, i.e., we do not plan a direct adaption of GOLD. For standard-conformant tagsets, then, the linking with this ontology becomes trivial. Still, as these meta-schems suffer from the problems of standardisation approaches in general, we suggest a harmonisation between our EAGLES-based ontology and GOLD. Accordingly, the terms used in EAGLES are provided with a formal definition retrievable from the mapping between EAGLES and GOLD.[21] Finally, other non-EAGLES conformant tagsets (SFB632 annotation standard, Tibetan tagset, SFB538/E2 tagset) will be integrated into the ontology.

Thus, our terminological backbone will be created in a three-step methodology:

1. derive an ontology from EAGLES,
2. harmonise this ontology with GOLD, and finally
3. integrate other non-EAGLES conformant tagsets.

The consideration of meta-tagsets instead of language-specific tagsets has several advantages:

- Considering meta tagsets rather than language-specific tagsets, the number of mapping procedures to integrate the tagsets applied to our data is reduced. As the „default mapping" between standard-conformant tagsets and meta-tagsets can be exploited (cf. Table 2), we have to consider four mappings (one from EAGLES/MULTEXT, one from the SFB632 tagset, one from the SFB538/E2 tagset, one from the Tibetan tagset). Otherwise, nine mappings would have to be implemented.
- By providing a default mapping between GOLD and EAGLES/MULTEXT, the effort to link other standard-conformant tagsets and GOLD is decreased. As a by-product, a natural extension of GOLD to a broad variety of languages even beyond the scope of the languages considered in the SFBs becomes likely, including the majority of Indo-European (Celtic, Germanic, Romance, Greek, Urdu, Slavonic), Basque and several Finno-Ugric (Estonian, Hungarian) languages.

---

[20] Indeed, there is another feature value "both" for entities which cannot be unambiguously considered pronouns or determiners. However, it is questionable whether such cases of surface-based ambiguity should be represented within the "terminological backbone" at all.

[21] As a consequence of the extension of EAGLES with formal definitions modifications of the "trivial" default mapping between standard-conformant tagsets and the meta-scheme might be necessary.

| meta tagsets and multilingual tagsets | | language-specific tagsets | languages |
|---|---|---|---|
| n/a | | Tibetan tagset | Classical/Old Tibetan, Ladakh, Balti |
| EAGLES | generalization of existing tagsets for European languages | SUSANNE | English |
| | | STTS, 3 variants | German |
| | | MENOTA | Old Norse |
| MULTEXT-East | (adaption of EAGLES for Eastern Europe) | Russian tagset | Russian |
| SFB632 annotation standard | designed for typological research | n/a | 13 typologically different languages |
| SFB538/E2 tagset | reduced tagset for acquisition studies | n/a | German, Romance, Basque |

**Table 2. Tagsets and meta-tagsets in the SFBs.**

## 5.4 Mapping tags to concepts

A s POS tags are labels for classes of words, we propose to integrate these directly into an ontology of word classes. This ontology consists of two principal components, the upper model which serves as a terminological backbone, and several domain models which correspond to individual tagsets.

The derivation of the upper model as a generalisation over different tag sets and meta tagsets was mentioned in the previous section. As an illustration, we consider the special case of nouns. The original definition in the EAGLES recommendations (Leech/Wilson 1996) is given as:

**Nouns (N)**
  (i)   Type:    1. Common     2. Proper
  (ii)  Gender: 1. Masculine  2. Feminine 3. Neuter
  (iii) Number: 1. Singular    2. Plural
  (iv)  Case:    1. Nominative 2. Genitive 3. Dative   4. Accusative  5. Vocative

Concentrating on the "Type" feature as a major subclassification among two distinctive parts of speech, we can derive a rudimentary taxonomy of nouns with the concept NOUN and two sub-concepts COMMONNOUN and PROPERNOUN. (Similar instances of "implicit" hierarchical structures are found throughout any tagset.)
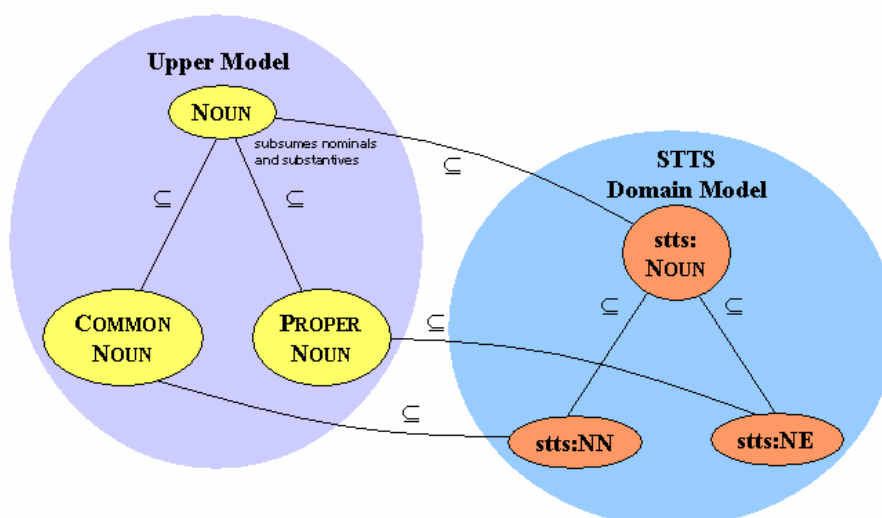


**Figure 2**. **Upper and domain model and their linking in the ontology. The case of nominals in STTS.**

These categories then have to be aligned with the corresponding categories in GOLD. The concept NOUN probably corresponds to NOUN$_{GOLD}$:[22] "a broad classification of parts of speech which include substantives and nominals", PROPERNOUN which is reserved explicitly for names seems to cover a sub-class of SUBSTANTIVE$_{GOLD}$ ("A substantive is a member of the syntactic class in which the names of physical, concrete, relatively unchanging experiences are most typically found [...]"), whereas COMMONNOUN possibly subsumes NOMINAL$_{GOLD}$ ("whose members differ gramatically from a substantive but which functions as one", such as GERUND$_{GOLD}$ which is "derived from a verb and used as a noun"), but certain instances of SUBSTANTIVE$_{GOLD}$ as well. At this point, we are already facing a problem with regard to GOLD, as its definitions are occasionally based on other conceptualisations than those applied in traditional Latin-based grammars underlying most European tagsets. Accordingly, impulses for possible modifications of the current version of GOLD are to be expected as another by-product of our research.

In a similar way, considering the German tagset STTS (Skut et al. 1998) as an example, a hierarchically structured *domain model* can be derived. Unlike the EAGLES recommendations, the STTS tagset is well-documented. The guidelines[23] give detailed enumerations of use-cases, examples of the categories and include enumerations of critical cases. Further, the aspect of hierarchical structuring is explicitly emphasised. So, the EAGLES-based upper model concepts NOUN and the sub-concepts COMMONNOUN and PROPERNOUN can be aligned easily with the (partial) tags N (subsuming NN and NE), NN (concrete and abstract nouns, measurements, ..., nominalised adjective, nominalised participle, etc.) and NE (names, surnames, trademarks, placenames, mountains, lakes, countries, etc.).

The linking between domain models and the upper model is implemented by means of conceptual subsumption (written $\subseteq$), resulting in a complex ontological structure, see Figure 2. To avoid confusion between concepts of the upper model and concepts of the domain models within the resulting ontology, namespaces are introduced. Technically, our ontology will be based on OWL/RDF.

## 5.5. Hybrid Concepts

One of the advantages of our ontology approach is the shift of complexity from the terminological backbone (the upper model in our terminology) to the linking between domain model and upper model. As a result, the upper model can be defined in a language-independent way, without the need to represent language-specific phenomena such as mergers between grammatical categories, surface-based ambiguity or fused forms. In the surface-oriented standardisation approach, however, hybrid concepts have been integrated to account for such phenomena.

Principally, the need for hybrid concepts arises from two sources: ambiguity and fusion. By fusion we mean the systematic contraction of different parts of speech into one compound form. As an example, there is a very common phenomenon of a fused preposition and an article in Western European languages. Accordingly, tagsets provide specialised tags for this category, e.g., APPRART in STTS. Thus, an additional value of adpositions had to be integrated into the EAGLES recommendations to account for such tagsets. Alternatively, this phenomenon "should preferably, however, be handled by assigning two tags to the same orthographic word (one for the preposition and one for the article)." (Leech/Wilson 1996). Similar to this preference, we suggest to model cases of fusion by the intersection of two upper model concepts, as illustrated in Figure 3.

Another source of hybrid forms is ambiguity. In a broad sense, several phenomena can be subsumed under ambiguity:

---

[22] The quotations from GOLD are taken from the HTML version of GOLD 0.2 available at www.linguistics-ontology.org/html-view (May 7th, 2006). To avoid ambiguities about the concepts, GOLD concepts will be written with the subscript $_{GOLD}$.
[23] The STTS guidelines (in German) can be found at http://www.sfs.uni-tuebingen.de/Elwis/stts.

- The same form can represent different underlying part of speech types. Accordingly, language-specific tagsets often introduce special tags to refer to such ambiguous words. As an example, the SUSANNE tag II is defined as "preposition, including prepositional use of word that can function either as preposition or as adverb" (Sampson 1996, p. 109)
- A related phenomenon is the systematic merging between different grammatical categories. Consider the English *–ing* forms: "An additional value to the non-finite category of verbs is arguably needed for English, because of the merger in that language of the gerund and participle functions. The –ing form does service for both and the two traditional categories are not easily distinguishable." (Leech and Wilson 1996)

To express the fact that the corresponding tag applies to words that belong to one of the categories between which ambiguity holds or merging occurred, we suggest referring to the union of two ambiguous concepts rather than making a choice between one of the possibilities. For the aforementioned problem of auxiliary verbs in STTS and the corresponding representation in our ontology, cf. Figure 4.
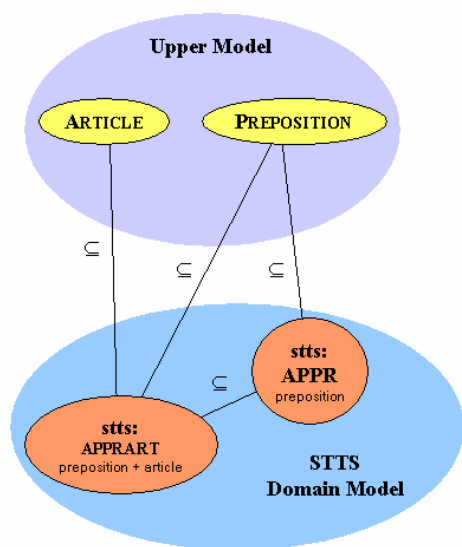


**Figure 3.** Representing fused forms in the linking.     **Figure 4.** Representing ambiguity in the linking.

$$\text{stts;APPRART} \subseteq \text{ADPOSITION} \cap \text{ARTICLE} \qquad \text{stts:VA} \subseteq \text{AUXILIARVERB} \cup \text{LEXICALVERB}$$

As a result of this shift of complexity from the "standard" onto the mapping, scalability and universality of the upper model will be enhanced.

# 6. Rules of Best Practice

In our experience, transforming data resources from project-specific formats to a more sustainable form has proven a difficult and time-consuming task, not least because many researchers are neither fully aware of potential sustainability problems, nor of existing solutions for avoiding them. If decisions made early in the process of designing a corpus took into account some rather broad recommendations for sustainable data handling, many problems arising later in the work with the corpus could probably be avoided altogether. Making such recommendations available to the research community in the form of Best Practice Guidelines can thus be crucial to the sustainability of future resources. Bird/Simons (2002) formulated a comprehensive set of such recommendations in their "Seven Dimensions of Portability for Language Documentation and Description". Virtually all the issues addressed in this paper are highly relevant to the work of our joint initiative, and we consider the corresponding recommendations appropriate and important for achieving our goals. Consequently, we expect to be able to contribute to a further development of these guidelines on several levels:

## 6.1 Data Creation and Documentation

Firstly, most of the issues addressed by Bird/Simons (2002) are meant to be valid for handling linguistic data in general and are therefore formulated neutrally, without targeting a specific audience of linguists – if an occasional focus can be discerned, it is on the community concerned with the documentation of endangered languages. We believe that it may be worthwhile to revisit the proposed seven dimensions from the perspective of other research communities represented at the three research centres. Reformulating the recommendations in terms of, for instance, child language acquisition research or conversation analysis, and illustrating good practices with concrete examples that are well-known and directly relevant to those researchers may considerably enhance the dissemination and acceptance of such guidelines in their respective communities[24].

Secondly, we expect our effort in the other work packages to reveal additional issues that can serve to substantiate and specify the recommendations suggested by Bird/Simons (2002). As an example, consider the development of data formats sketched in section 4.1 above. The relevant recommendations by Bird/Simons (2002) are the ones in the dimension "Format" which comprises the four subareas "Openness", "Encoding", "Markup" and "Rendering". They basically point the reader in the direction of using XML and Unicode (or at least similar open standards), but otherwise make no detailed mention of existing XML-based data models or give recommendations on how to decide between competing models. As has been pointed out above, mediating between different approaches to XML-based handling of linguistic data is a central issue in our joint initiative. We expect this task to reveal some more general criteria for deciding, for example, whether and where hierarchy-based data models are preferable to time-based ones (or vice versa), or under what circumstances standoff approaches to annotation are (or are not) in the interest of sustainability. These general criteria may then complement the recommendations given in Bird/Simons (2002) in the "Format" dimension. The same holds for other dimensions and other work packages of our initiative.

Thirdly, some "dimensions of portability" may require a specification simply because they highly depend on the specific context in which a resource is created and used. This is especially true of the Dimensions "Rights" and "Citation" where national law or the citation

---

[24] More trivially, providing a German translation of the guidelines may also be important in that respect.

practices of a specific research community may predetermine to a great part what can be recommended as a best practice.[25] The next section will discuss this in more detail.

## 6.2 Legal Questions in Data Archiving

The major part of the three research centres' project data has been collected and will be reused primarily in environments in which German and/or European law is applicable. For this reason the main focus of the current work is on legal issues, which result from national law. Legal questions dealing with the international exchange and reuse of research data will be part of future work, however the principles described below do not depend only on national law. Corpus content, annotation schemes, and access software to be used for scientific purposes can be classified into two different types of data that are subject to different aspects of legal protection:

- Immaterial Goods – Non-material goods which are any kind of intellectual property of a third party, such as copyrighted work. This includes databases, software, and utility patents etc. Applicable law in most of these cases is the respective national Copyright Act (in Germany the *Urheberrechtsgesetz – UrhG*)
- Personal Data – Data that are linked to an individual, e.g., audio and video recordings, any speech transcriptions as well as metadata that contain personal information on speakers. Applicable law for the reuse and exchange of this kind of data is the respective data protection act (e.g., the German *Bundesdatenschutzgesetz – BDSG*)

Some of the main issues to deal with are caused by the German UrhG[26] that protects an author's intellectual property rights for *literary*, *scientific* and *artistic works* (see Art 1 UrhG[27]). Like the US American Copyright, it expires 70 years after an author's death, but in contrast to US American law the German UrhG does not allow an author to assign copyright to a third party. Against this background any reuse and exchange of project data written by a third party author would infringe copyright. In fact, in most cases an uncertain legal situation is one of the major arguments produced against this.

In terms of law, language corpora themselves are defined as *databases*. The legal basis for the protection of database works in Europe is provided by the *Directive 96/9/EC of the European Parliament on the legal protection of databases*[28] that defines a database as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means". Because building a database requires the investment of extensive technical and human as well as financial resources, the copyright on databases is divided into *authors* and *makers* copyright: whereas a database author is defined as the one who created the structure and layout of a database, the maker is the one who has made the investment. This distinction is highly relevant to third party funded research that produces databases, such as the three research centres involved in this project.

Further potential subjects of legal protection are computer programs to be used for accessing a database. Such programs are not protected by the EU-Directive and must be handled as a separate subject of copyright protection.

To summarise, there are four different subgroups of potential holders of copyrights on linguistic corpora:

---

[25] See Baude et al. (2005) for an existing best practice guideline that is much more specific and detailed in that respect because it takes into account many factors that are of special relevance to researchers working with corpora in France.
[26] See http://www.gesetze-im-internet.de/urhg/index.html
[27] English translation of the UrhG by IUSCOMP (http://www.iuscomp.org/)
[28] http://europa.eu.int/ISPO/legal/en/ipr/database/database.html

1. authors of corpus content
2. authors of corpus databases
3. authors of database access software
4. makers of the database (investors)

As mentioned above, the reuse and exchange of linguistic data may not only refer to aspects of copyright, but to data protection law. This case is given by personal data: "any information concerning the personal or material circumstances of an identified or identifiable individual (the data subject)".[29] According to this, there is a fifth group of persons, whose rights concerning the data can be asserted:

5. subjects and test persons

Common practice when dealing with personal data is *anonymisation* (removing personal information on individuals by abbreviation of names, locations etc.) or *synonymisation* (renaming individuals, locations, etc.). In some cases, however, personal information may be relevant for linguistic analysis, in other cases (especially audio and video recordings) a full anonymisation would require considerable technical effort and might cause damage to the data.

To deal with these problems, anonymisation and synonymisation need to be integrated into the annotation scheme, so that personal information in transcriptions and metadata can be reconstructed if legislation does allow this. Data that cannot be anonymised appropriately might be excluded from distribution, or the distribution will be bound to contractual agreements with the subjects.

In addition to the evaluation of the (European and international) legal positions of legal protection, working out these contractual agreements to be concluded with subjects and authors will be one of the central future tasks.

## 7. Conclusion and Future Work

This paper has outlined a new research initiative aiming at solutions for sustainable archiving of linguistic data. It has presented previous work in which diverse bodies of data in project specific formats were integrated into less restricted XML-based frameworks, and it has identified seven areas of open questions on the way from such frameworks to truly sustainable archives.

If there is a conclusion to be drawn at this early stage of our joint initiative, it is that we do not have to start from zero. The spread of XML as a widely used and supported standard for representing language data provides a much more solid basis for our work than was available six or seven years ago. The fact that our existing solutions are already based on that standard and are also related to ongoing work on more general text-technological frameworks for linguistic data processing also makes the task of finding a suitable common data format more easily definable. In other areas, too, we can profit from work of the language resource community done in the last decade. Thus, metadata standards like IMDI and OLAC are bound to play an important role in our sustainability effort. Likewise, existing best practice guidelines like the ones proposed by Bird/Simons (2002) provide a good starting point for our own work in that area. Last but not least, the difficult task of terminology integration can profit considerably from the work done on the GOLD ontology.

---

[29] BDSG, Section 1 Art 3 I, English translation by http://www.datenschutz-berlin.de/recht/de/bdsg/bdsg01_eng.htm

## Acknowledgments

## References

Bayerl, Petras S., Harald Lüngen, Daniela Goecke, Andreas Witt, and Daniel Naber2003. Methods for the semantic analysis of document markup. In *Proceedings of the 2003 ACM Symposium on Document Engineering* (Grenoble, France, November 20 - 22, 2003). DocEng '03. ACM Press, New York, NY, 161-170. (http://doi.acm.org/10.1145/958220.958250)

Baude, Olivier, Claire Blanche-Benveniste, Marie-France Calas, Pascal Cordereix, Isabelle De Lambertierie, Laurence Goury, Michel Jacobson, Christiane Marchello-Nizia, and Lorenza Mondada (2005): Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux, Centre National de la Recherche Scientifique, Paris.

Bird, Steven and Liberman, Mark (2001): A formal framework for linguistic annotation. In: Speech Communication 33(1,2), 23-60.

Bird, Steven and Simons, Gary (2002): Seven Dimensions of Portability for Language Documentation and Description. In: Language 79, 557-582.

Brants, T., S. Dipper, S. Hansen, W. Lezius, G. Smith (2002), The TIGER Treebank. In: Proc. of the Workshop on Treebanks and Linguistic Theories. Sozopol.

Broschart, J. (1997). Why Tongan does it differently: Categorial distinctions in a language without nouns and verbs. *Linguistic Typology* 1-2, 123-166.

Carletta, Jean, Jonathan Kilgour, Timothy O'Donnell, Stefan Evert, 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).

Chiarcos, Ch. and Krasavina, O. (2005), PoCoS – Potsdam Coreference Scheme, version of June 2005, http://amor.cms.hu-berlin.de/~krasavio/annorichtlinien.pdf (May 7th, 2006)

Dipper, Stefanie, Michael Götze, Manfred Stede and Tillmann Wegst (2004): ANNIS: A Linguistic Database for Exploring Information Structure In Ishihara, S., M. Schmitz and A. Schwarz (eds.): Interdisciplinary Studies on Information Structure 1 (ISIS), pp. 245-279, Potsdam, Germany, http://www.ling.uni-potsdam.de/~dipper/papers/isis04.pdf (May 9th, '06)

Dipper, Stefanie, Michael Götze, Manfred Stede (2005). Heterogeneity in Focus: Creating and Using Linguistic Databases. In Ishihara, S., Schmitz, M. (Eds.), Interdisciplinary Studies on Information Structure (ISIS), 2. Potsdam: Universitätsverlag Potsdam.

Dipper, Stefanie, Erhard Hinrichs, Thomas Schmidt, Andreas Wagner, and Andreas Witt (2006): Sustainability of linguistic resources. In: Proceedings of the LREC Workshop on merging and layering linguistic information. Genoa.

EAGLES 2000. Expert Advisory Group on Language Engineering Standards (EAGLES). 2000. XCES Corpus Encoding Standard for XML. XML version of the CES DTDs. Document XCES 0.2. 10 February 2000.

Erjavec, Tomaž (2004): MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris.

Farrar, Scott and D. Terence Langendoen (2003). A Linguistic Ontology for the Semantic Web. GLOT International 7(3), pp. 97-100.

Féry, C.; Skopeteas, St. and Stoel, R. (2006), Typological Data on Information Structure. Paper presented at Int. Conf. on Linguistic Evidence (LingEvid) 2006, February 2006, Tübingen.

Hinterhölzl, R. and Petrova, S. (2005). Rhetorical Relations and Verb Placement in Early Germanic Languages. Evidence from the Old High German Tatian translation (9th century). In Stede, M. et al. (Eds.), *Salience in Discourse. Multidisciplinary Approaches to Discourse*, 71-79.

Hughes, J.S.; Souter, C. and Atwell, E.S. (1995), Automatic extraction of tagset mappings from parallel annotated corpora. In: Tzoukermann E. and Armstrong, S. (editors), From Text to Tags: Issues in Multilingual Language Analysis, Proc. ACL-SIGDAT Workshop, pp.10-17.

Ide, Nancy, Laurent Romary, Eric de la Clergeri, (2005). International Standard for a linguistic annotation framework. In Proceedings of HLT-NAACL'03 Workshop on the Software Engineering and Architecture of Language Technology.

Khoja, S, Garside, R, and Knowles, G (2001) A tagset for the morphosyntactic tagging of Arabic. Paper given at the Corpus Linguistics 2001 conference, Lancaster.

Leech, G. and Wilson, A. (1996), EAGLES Recommendations for the Morphosyntactic Annotation of Corpora, Version of Mar, 1996, http://www.ilc.cnr.it/EAGLES/annotate/annotate.html, May 4th, 2006.

Müller, Ch. and M. Strube (2001), MMAX: A tool for the annotation of multi-modal corpora. In: Proceedings if the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems. Seattle, August 2001, pp. 45-50.

Nowak, E. (2005), Lexical Categories in Polysynthetic Languages. In: Cruse, A. et al. (eds.), *Handbuch Lexikologie – Lexicology.* Berlin: de Gruyter, 981-986.

O'Donnell, M. (1997), RST-Tool: An RST analysis tool. In: Proc. of the 6th European Workshop on Natural Language Generation, Duisburg, 1997.

Rehbein, Jochen / Schmidt, Thomas / Meyer, Bernd / Watzke, Franziska / Herkenrath, Annette (2004): Handbuch für das computergestützte Transkribieren nach HIAT. Arbeiten zur Mehrsprachigkeit Folge B (Nr. 56). Universität Hamburg: Sonderforschungsbereich Mehrsprachigkeit.

Reiter, D. and M. Stede (2003), Step by step: Underspecified markup in incremental rhetorical analysis. In: Proc. of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), Budapest.

Sampson, G. (1995), English for the Computer. Oxford: Clarendon Press.

Schmidt, Thomas (2005a): Computergestützte Transkription - Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln (Sprache, Sprechen und Computer/Computer Studies in Language and Speech 7). Frankfurt a. M.

Schmidt, Thomas (2005b): Time-based data models and the Text Encoding Initiative's

Skut, W., Brants, T., Krenn, B. & Uszkoreit, H. (1998). A Linguistically Interpreted Corpus of German Newspaper Text. ESSLI-1998, Workshop on Recent Advances in Corpus Annotation.

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

Stede, M. (2004), The Potsdam Commentary Corpus. In *Proceedings of the ACL-04 Workshop on Discourse Annotation*. Barcelona, Spain.
Wagner (2005): "Unity in Diversity", In Dipper et al. (2005), pp. 1 – 20

Wagner and Zeisler (2004). A syntactically annotated corpus of Tibetan. In Proc. of LREC 2004, pages 1141–1144, Lisboa, May 2004.

Witt, A. (2004), Multiple Hierarchies: New aspects of an old solution. Proceedings of Extreme Markup Languages 2004. Montreal, Canada.