Erschienen in: Storrer, Angelika (Hrsg.): Proceedings of KONVENS 2008. KONVENS 2008 - Ergänzungsband Berlin: BBAW, 2008. S. 86-96.

SPLICR: a sustainability platform for linguistic corpora and resources

Georg Rehm, Oliver Schonefeld, Andreas Witt,
Christian Chiarcos, Timm Lehmberg

Abstract. We present SPLICR, the Web-based Sustainability Platform for Linguistic Corpora and Resources. The system is aimed at people who work in Linguistics or Computational Linguistics: a comprehensive database of metadata records can be explored in order to find language resources that could be appropriate for one's specific research needs. SPLICR also provides an interface that enables users to query and to visualise corpora. The project in which the system is being developed aims at sustainably archiving the ca. 60 language resources that have been constructed in three collaborative research centres. Our project has two primary goals: (a) To process and to archive sustainably the resources so that they are still available to the research community in five, ten, or even 20 years time. (b) To enable researchers to query the resources both on the level of their metadata as well as on the level of linguistic annotations. In more general terms, our goal is to enable solutions that leverage the interoperability, reusability, and sustainability of heterogeneous collections of language resources.

1 Introduction

This contribution presents SPLICR, the Web-based Sustainability Platform for Linguistic Corpora and Resources aimed at people who work in linguistics or computational linguistics: a comprehensive database of metadata records can be explored and searched in order to find language resources that could be appropriate for one's specific research needs. SPLICR also provides a graphical interface that enables users to query and to visualise corpora.

The project in which SPLICR is being developed aims at sustainably archiving (Trilsbeek and Wittenburg 2006) the language resources that have been constructed or are still work in progress in three collaborative research centres (Sonderforschungsbereiche). The groups in Tübingen (SFB 441: "Linguistic Data Structures"), Hamburg (SFB 538: "Multilingualism"), and Potsdam/Berlin (SFB 632: "Information Structure") built a total of 56 resources – corpora and treebanks mostly. According to estimates it took more than one hundred person years to collect and to annotate these datasets. Our project has two goals: (a) To process and to sustainably archive the resources so that they are still available to the research community and other interested parties in five, ten, or even 20 years time (Schmidt et al. 2006). (b) To enable researchers to query the resources both on the level of their metadata as well as on the level of linguistic annotations. In more general terms, our main goal is to enable solutions that leverage the interoperability, reusability, and sustainability of a large collection of heterogeneous language resources.

The main advantage of the system and its underlying architecture is that we designed and specified an integrated workflow that starts with the processing of individual corpora at multiple sites using custom-made tools. Afterwards, the processed corpora along with metadata files, their original data sets, HTML- or PDF-based manuals and transformation logfiles are copied into a directory tree whose structure is specified by rigid protocols. In the next step, this directory tree is traversed using a lightweight importer client that checks the directory tree for consistency and copies the files of the heterogeneous resources onto the SPLICR server that, in turn, provides a homogeneous web-based means of access.

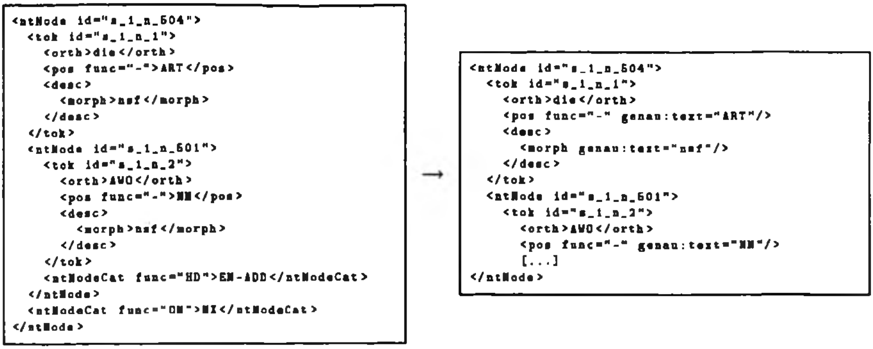
The remainder of this paper is structured as follows: section 2 introduces our approach to normalising corpus data (section 2.1) and metadata records (section 2.2). SPLICR's architecture is described in section 3, although we are only able to highlight selected parts of the system due to space restrictions. The staging area is briefly discussed in section 3.1, while section 3.2 gives an overview of our approach to representing knowledge about linguistic annotation schemes using ontologies. A third major component of the system is the graphical corpus query and visualisation frontend (section 3.3). The article ends with concluding remarks (section 4).

2 Data normalisation and representation

One of the obstacles we are confronted with is providing homogeneous means of accessing a large collection of diverse and complex linguistic resources. For this purpose we developed several custom tools in order to normalise the corpora (section 2.1) and their metadata records (section 2.2).

2.1 Normalisation of linguistic resources

Language resources are usually built using XML-based languages nowadays (Ide et al. 2000; Lehmberg and Wörner 2007; Sperberg-McQueen and Burnard 2002; Wörner et al. 2006) and contain several concurrent annotation layers that correspond to multiple levels of linguistic description (e. g. part-of-speech, syntax, coreference). Our approach includes the normalisation of XML-annotated resources, e. g. for cases in which corpora use PCDATA content to capture both primary data (i. e. the original text or transcription) as well as annotation information (e. g. POS tags). We use a set of tools to ensure that only primary data is encoded in PCDATA content and that all annotations proper are encoded using XML elements and attributes. The



```
<ntNode id="s_1_n_504">
  <tok id="s_1_n_1">
    <orth>die</orth>
    <pos func="-">ART</pos>
    <desc>
      <morph>nsf</morph>
    </desc>
  </tok>
  <ntNode id="s_1_n_601">
    <tok id="s_1_n_2">
      <orth>AWO</orth>
      <pos func="-">NN</pos>
      <desc>
        <morph>nsf</morph>
      </desc>
    </tok>
    <ntNodeCat func="HD">EN-ADD</ntNodeCat>
  </ntNode>
  <ntNodeCat func="ON">NX</ntNodeCat>
</ntNode>
```

→

```
<ntNode id="s_1_n_504">
  <tok id="s_1_n_1">
    <orth>die</orth>
    <pos func="-" genau:text="ART"/>
    <desc>
      <morph genau:text="nsf"/>
    </desc>
  </tok>
  <ntNode id="s_1_n_601">
    <tok id="s_1_n_2">
      <orth>AWO</orth>
      <pos func="-" genau:text="NN"/>
      [...]
</ntNode>
```

Figure 1. An example from the TüBa-D/Z treebank (represented in the Tusnelda format) before (left) and after processing the resource with our normalisation tools (right)

transformation from PCDATA content (i. e. XML elements) to CDATA values (i. e. XML attributes) is performed semi-automatically.

Figure 1 illustrates this process by means of an excerpt from the TüBa-D/Z treebank (Telljohann et al. 2004) in one of its four representation formats (Tusnelda, see Wagner 2005). Beside the actual primary data content "die AWO" (the PCDATA content of the XML element <orth>) other XML elements such as <pos> use PCDATA content to encode grammatical information. Since this information serves annotation purposes, the contents of elements that do not contain primary data within their PCDATA content are transformed to the value of the attribute genau:text that is introduced by our tools. As Tusnelda documents comprise several levels of annotation in a single, monolithic XML element tree, the overall annotation is still extremely complex even though we perform a normalisation procedure that includes the step sketched above. Therefore, we use additional processing methods to split the different conceptual levels, e. g. syntax, morphology, and named entities into multiple documents, that is, into a multi-rooted tree (Witt et al. 2007).

Another reason for the normalisation procedure is that both hierarchical and timeline-based corpora (Bird and Liberman 2001; Schmidt 2005) need to be transformed into a common annotation approach, because we want our users to be able to query both types of resources at the same time and in a uniform way. Our approach (Dipper et al. 2006; Schmidt et al. 2006; Wörner et al. 2006) can be compared to the NITE Object Model (Carletta et al. 2003): we developed tools that semiautomatically split hierarchically annotated corpora that typically consist of a single XML document instance into individual files, so that each file represents the information related to a single annotation layer (Rehm et al. 2008b; Witt et al. 2007); this approach also guarantees that overlapping structures can be represented straightforwardly. Timeline-based corpora are also processed in order to separate graph annotations. This approach enables us to represent arbitrary types of XML-annotated corpora as individual files, i. e. individual XML element trees. These are encoded as regular XML document instances, but, as a single corpus comprises *multiple* files, there is a need to go beyond the functionality offered by typical XML tools to enable us to process multiple files, as regular tools work with single files only (our approach for querying multi-rooted trees is described by Rehm et al. 2007a, 2008a).

Almost all resources that we process are linguistic corpora and treebanks. In addition, there are a few resources that belong to different data types. Four SFB 441 projects construct sentence collections that consist of, for example, suboptimal syntactic constructions taken from the linguistic literature and annotated with grammaticality judgements, or sentences that have specific verb phrases such as the stative passive. Furthermore, multiple projects create lexicons, some of which are augmented with empirical judgements gathered in online experiments. In a secondary line of research we develop generic XML-based representation formats for these types of linguistic resources for which we also implement query and visualisation methods.

2.2 Normalisation of metadata records

The separation of the individual annotation layers contained in a corpus has serious consequences with regard to legal issues (Lehmberg et al. 2007a,b, 2008; Rehm et al. 2007c; Zimmermann and Lehmberg 2007): due to copyright and personal rights specifics that usually apply to a corpus's primary data we provide a fine-grained access control layer to regulate access by means of user accounts and access roles. We have to be able to explicitly specify that a certain user only has access to the set of, say, six annotation layers (in this example they might be available free of charge for research purposes) but not to the primary data, because they might be copyright-protected (Rehm et al. 2007c,d).

Our generic metadata schema, eTEI, is based on the TEI P4 header (Sperberg-McQueen and Burnard 2002) and extended by a set of additional requirements. Both eTEI records and the corpora are stored in an XML database. The underlying assumption is that XML-annotated datasets are more sustainable than, for example, data stored in a proprietary relational DBMS. The main difference between eTEI and other approaches is that the generic eTEI metadata schema, currently formalised as a single document type definition (DTD), can be applied to five different levels of description (Himmelmann 2006; Trippel 2004). One eTEI file contains information on one of the following levels: (1) *setting* (used for recordings or transcripts of spoken language, describes the situation in which the speech or dialogue took place); (2) *raw data* (e. g. a book, a piece of paper, an audio or video recording of a conversation etc.); (3) *primary data* (transcribed speech, digital texts etc.); (4) *annotations*;

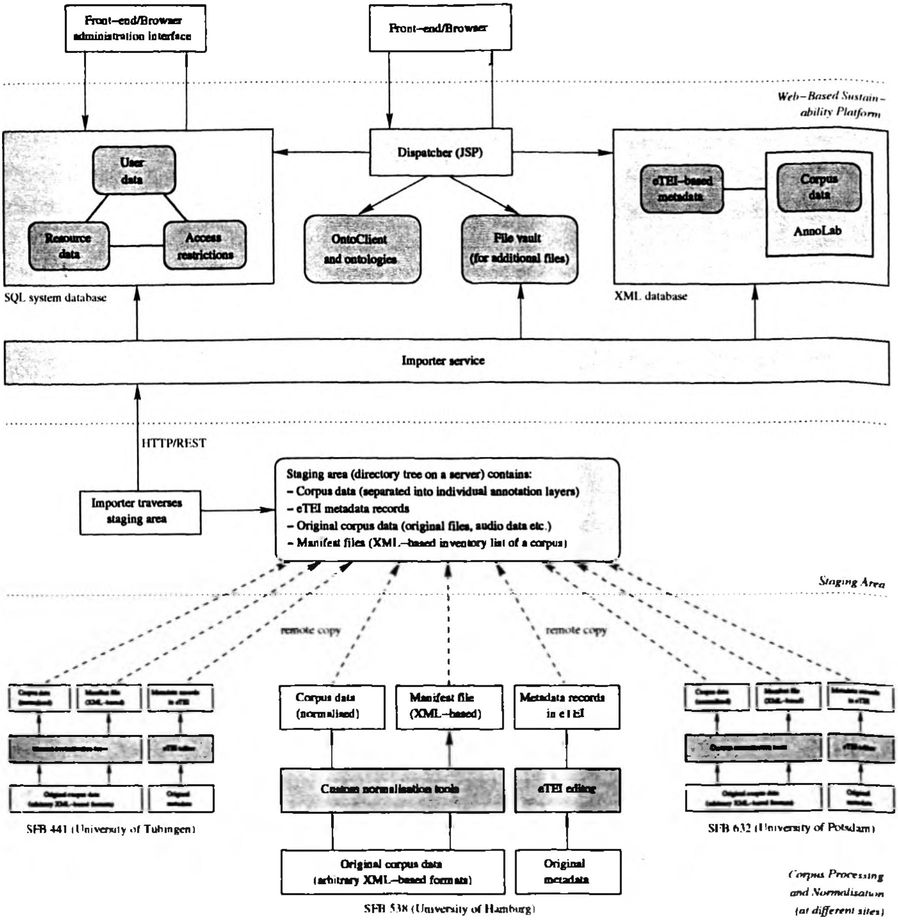

Front–end/Browser administration interface
Front–end/Browser
Web–Based Sustain–ability Platform
User data
Resource data
Access restrictions
Dispatcher (JSP)
OntoClient and ontologies
File vault (for additional files)
eTEI–based metadata
Corpus data
AnnoLab
SQL system database
XML database
Importer service
HTTP/REST
Importer traverses staging area
Staging area (directory tree on a server) contains:
– Corpus data (separated into individual annotation layers)
– eTEI metadata records
– Original corpus data (original files, audio data etc.)
– Manifest files (XML–based inventory list of a corpus)
Staging Area
remote copy
remote copy
Corpus data (normalised)
Manifest file (XML–based)
Metadata records in eTEI
Custom normalisation tools
eTEI editing
Original corpus data (arbitrary XML–based formats)
Original metadata
SFB 441 (University of Tubingen)
SFB 538 (University of Hamburg)
SFB 632 (University of Potsdam)
Corpus Processing and Normalisation (at different sites)

Figure 2. Resource normalisation, the staging area and the primary SPLICR components

(5) *a corpus* (consists of primary data with one or more annotation levels). The need for these five levels of metadata description can be illustrated using the ambiguity of the "author" concept: while *setting* refers to a specific communication situation, the author of *raw data* can be the author of a certain book or the speaker whose monologue has been recorded. The author of *primary data* is the person who transcribed the raw data into a set of digital files. The authors of individual *annotation* files are those who analyse and interpret the primary data (usually linguists, student assistants or PhD students) and the author of the *corpus* is the person who is responsible for constructing or collecting the corpus data (for example, the principal investigator of

a research project). Important metadata exist on all five levels and can be captured using the approach described in this section.

We devised a workflow that helps users edit eTEI records (Rehm et al. 2008b). Its primary components are the eTEI DTD and the Oxygen XML editor. Based on annotations contained in the DTD we can generate automatically an empty XML document with embedded documentation and a Schematron schema. The Schematron specification can be used to check whether all elements and attributes instantiated in an eTEI document conform to the current level of metadata description.

3 Architecture

The sustainability platform consists of a front-end and a back-end. The front-end is the user visible part and is realised using JSP (Java Server Pages) and Ajax technologies. It runs in the user's browser and provides functions for searching and exploring metadata records and corpus data. The back-end hosts the JSP files and related data. It accesses two different databases, the *corpus database* and the *system database*, as well as a set of ontologies and additional components.[1] The corpus database is an XML database, extended by the AnnoLab system (Eckart and Teich 2007), in which all resources and metadata are stored. The system database is a relational database that contains all data about user accounts, resources (i. e. annotation layers), resource groups (i. e. corpora) and access rights. A specific user can only access a specific resource if the permissions for this user/resource tuple allow it.

The following subsections describe three selected parts of SPLICR's architecture: the staging area (section 3.1), a set of ontologies of linguistic annotations (section 3.2) and the querying front-end (section 3.3).

3.1 Staging area

A new resource is imported into the sustainability platform by (remotely) copying all corresponding files into the staging area whose directory structure is defined in a technical specification. Strict naming rules apply for the processed files (see section 2) and for the directories so that the whole directory tree can be traversed and processed automatically. Each corpus contains a manifest file, that is represented in a simple XML format and that acts as a corpus inventory. Manifest files are automatically generated by the normalisation tools, their contents are used by the GUI and by

[1] In the file vault area, SPLICR contains additional data about a resource, such as the original corpus data files, PDF files that act as documentation, and transformation scripts, amongst others. These additional files are available through the user interface as well by providing access via HTTP.

the import and export tools. The importer traverses the staging area, checks, among others, the data for consistency and imports the corpus data and metadata records into the XML database (we currently use eXist but are exploring several alternatives) using a REST-style HTTP interface. At the same time, new resource and resource group records as well as permissions are set up in the system database (MySQL). Permissions are chosen based on the restrictions defined in metadata records.

3.2 Ontologies of linguistic annotation

The corpora that we process are marked up using several different markup languages and linguistic tag sets. As we want to enable users to query multiple corpora at the same time, we need to provide a unifying view of the markup languages used in the original resources. For this sustainable operationalisation of existing annotation schemes we employ the ontologies of linguistic annotation (OLiA) approach: we built an OWL DL ontology that serves as a terminological reference. This reference model is based on the EAGLES recommendations for morphosyntax, the general ontology for linguistic description (Farrar and Langendoen 2003), and the LISA annotation standard (Dipper et al. 2007). It covers reference specifications for word classes, and morpho-syntax (Chiarcos 2008), and is currently extended to syntax and information structure. The OLiA reference model represents a terminological backbone that different annotations are linked to and consists of three components: a taxonomy of linguistic categories (OWL classes such as NOUN, COMMONNOUN), a taxonomy of grammatical features (OWL classes, e. g. ACCUSATIVE), and relations (OWL properties, e. g. HASCASE). An OLiA annotation model is an ontology that represents one specific annotation scheme (see figure 3). We built, among others, annotation models for the LISA annotation format (Dipper et al. 2007) used in typological research, TIGER/STTS (Brants et al. 2003; Schiller et al. 1999), two tag sets for Russian and five tag sets for English, e. g. Susanne (Sampson 1995), and PTB (Marcus et al. 1993). The linking between annotation models and the reference model is specified in separate OWL files.

Any tag from an annotation model can be retrieved from the reference model by a description in terms of OWL classes and properties. For this task, OntoClient was developed, a query preprocessor implemented in Java that uses an OWL DL reasoner to retrieve the set of individuals that conform to a particular description with regard to the reference model. The OntoClient enables us to use abstract linguistic concepts such as `Noun and hasCase(Accusative)` in a query. By means of an XQuery extension function, these concepts are expanded into the concrete tag names used in the annotation schemes of the corpora that are currently in the user's focus. The expansion is thus guided by annotation metadata as described in section 2.2.

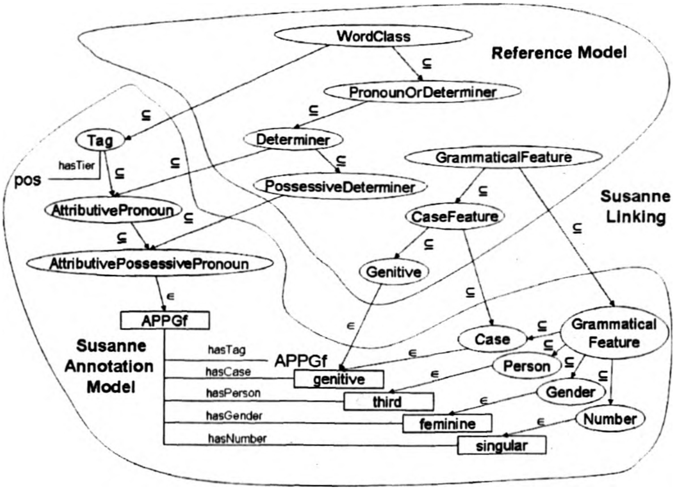
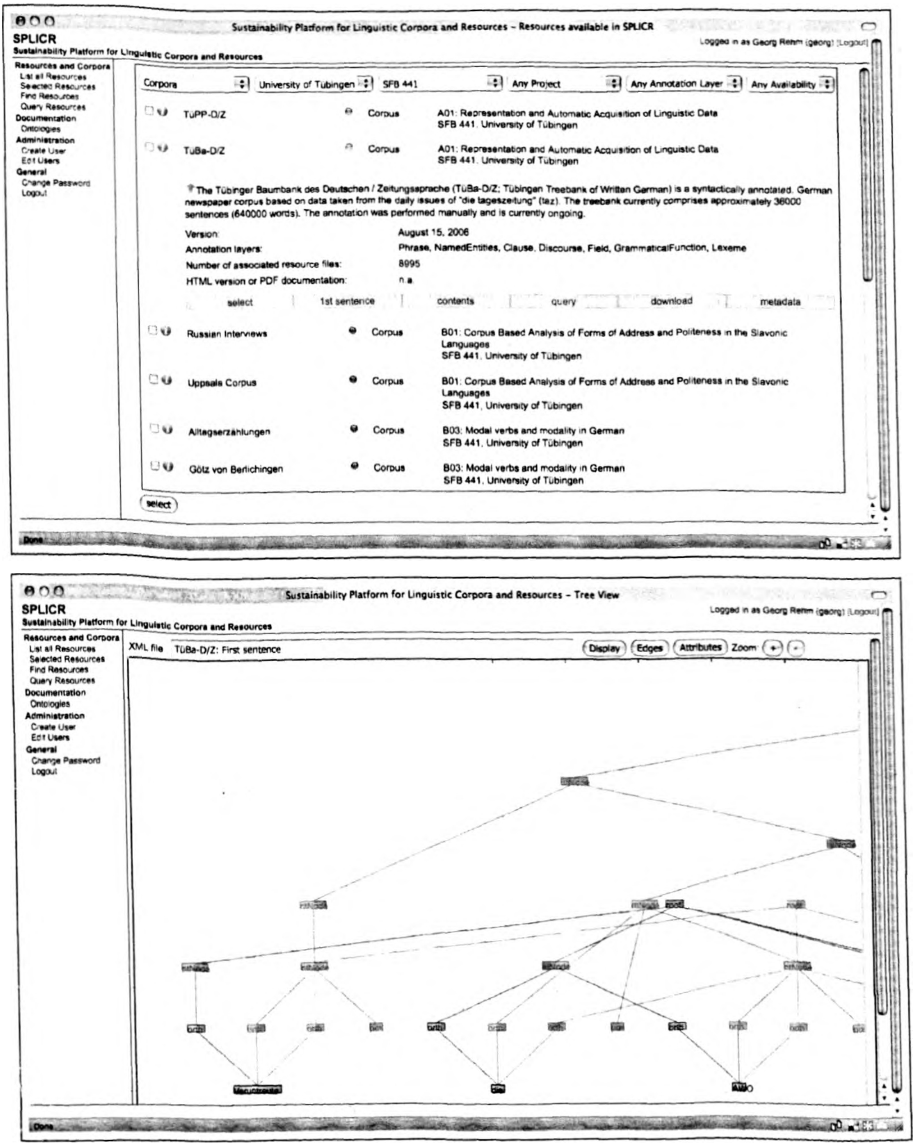


Figure 3. The Susanne tag **APPGf**, its representation within the annotation model and linking with the reference model

3.3 The corpus query front-end

As we cannot expect our target users (i. e. linguists) to be proficient in XML query languages such as XQuery, we provide an intuitive user interface that generalises from the underlying data structures and querying methods actually used. The ontology of linguistic annotations (section 3.2) provides abstract representations of linguistic concepts (e. g. *Noun*, *Verb*, *Preposition*) that may have a specific set of features; operands can be used to glue together the linguistic concepts by dragging and dropping these graphical representations onto a specific area of the screen, building a query step by step. We collected a set of requirements and functions that the front-end should have (such as the ones briefly sketched at the beginning of this section) by conducting in-depth interviews with the staff members of SFB 441 and by asking them to fill out a questionnaire (Soehn et al. 2008).

The front-end is implemented in JavaScript extended by the frameworks Prototype and script.aculo.us. One of its central components is a graphical tree fragment query editor that supports the processing of multi-layer annotations and that interprets and translates graphical queries into XQuery. The front-end communicates with the backend via Ajax, posting XQuery requests to a servlet running on the backend. The servlet responds with the XML-encoded matches, which are then interpreted by a variety of display modules. Five major display modes are already implemented: plain text view, XML view, box view, graphical tree view and timeline view.



□	TüPP-D/Z	Corpus	A01: Representation and Automatic Acquisition of Linguistic Data SFB 441, University of Tübingen
□	TüBa-D/Z	Corpus	A01: Representation and Automatic Acquisition of Linguistic Data SFB 441, University of Tübingen

The Tübinger Baumbank des Deutschen / Zeitungssprache (TüBa-D/Z; Tübingen Treebank of Written German) is a syntactically annotated, German newspaper corpus based on data taken from the daily issues of "die tageszeitung" (taz). The treebank currently comprises approximately 36000 sentences (640000 words). The annotation was performed manually and is currently ongoing.

Version:	August 15, 2006
Annotation layers:	Phrase, NamedEntities, Clause, Discourse, Field, GrammaticalFunction, Lexeme
Number of associated resource files:	8995
HTML version or PDF documentation:	n.a.

□	Russian Interviews	Corpus	B01: Corpus Based Analysis of Forms of Address and Politeness in the Slavonic Languages SFB 441, University of Tübingen
□	Uppsala Corpus	Corpus	B01: Corpus Based Analysis of Forms of Address and Politeness in the Slavonic Languages SFB 441, University of Tübingen
□	Alltagserzählungen	Corpus	B03: Modal verbs and modality in German SFB 441, University of Tübingen
□	Götz von Berlichingen	Corpus	B03: Modal verbs and modality in German SFB 441, University of Tübingen

XML file TüBa-D/Z: First sentence

Figure 4. The front-end in resource listing (above), and multi-rooted tree display mode (below)

The tree fragment query editor (Rehm et al. 2008a) involves dragging and dropping elements on an assembly pane, so that queries can be constructed in a step-by-step fashion. At the moment, structural nodes can be combined by dominance, precedence, and secondary edge relations. The structures defined by these graphs mirror the structures to be found. Each node may contain one or more conditions linked by Boolean connectives that help to refine the node classes allowed in the structures. We plan to realise a set of functions that can be roughly compared to TIGERSearch's feature set (Lezius 2002) enhanced by our specific requirements, i. e. multi-layer querying and query expansion through ontologies.

4 Concluding remarks and future work

The research presented in this contribution is still work in progress. We want to highlight some of the aspects that we plan to realise by the end of 2008. While the corpus normalisation and preprocessing phase is, with only minor exceptions, finished, the process of transforming the existing metadata records into the eTEI format was completed in June. Work on the querying engine and integration of the XML database, metadata exploration and on the graphical visualisation and querying front-end (Rehm et al. 2008a) as well as on the back-end is ongoing; we plan to finish work on the first prototype of the platform by September.

In addition we plan several extensions and modifications for the eTEI schema. Most notably, we plan to replace the current DTD, based on TEI P4, with an XML Schema description that is based on the current version of the guidelines (P5) and realised by means of an ODD ("one document does it all") specification. XML Schema has better and more appropriate facilities for including embedded documentation than the rather simple and unstructured comments available in DTDs. Another area that needs further work is the query front-end that we plan to upgrade and to enhance. In addition to a substantial overhaul of the interface in order to improve its usability, we will integrate query templates and saved searches that act like bookmarks in a web browser.

Bibliography

Bird, Steven and Mark Liberman (2001). A Formal Framework for Linguistic Annotation. *Speech Communication* 33(1/2):23–60.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit (2003). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation* 2(4):597–620.

Carletta, Jean, 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).*

Chiarcos, Christian (2008). An Ontology of Linguistic Annotations. *LDV Forum* 23(1):1–16.

Dipper, Stefanie, Michael Götze, and Stavros Skopeteas (eds.) (2007). *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, volume 7 of *ISIS*.

Dipper, Stefanie, Erhard Hinrichs, Thomas Schmidt, Andreas Wagner, and Andreas Witt (2006). Sustainability of Linguistic Resources. In Erhard Hinrichs, Nancy Ide, Martha Palmer, and James Pustejovsky (eds.), *Proceedings of the LREC 2006 Satellite Workshop Merging and Layering Linguistic Information*, 48–54, Genoa, Italy.

Eckart, Richard and Elke Teich (2007). An XML-Based Data Model for Flexible Representation and Query of Linguistically Interpreted Corpora. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer (eds.), *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, 327–336, Tübingen: Narr.

Farrar, Scott and Terry Langendoen (2003). A Linguistic Ontology for the Semantic Web. *GLOT International* 3:97–100.

Himmelmann, Nikolaus P. (2006). Daten und Datenhuberei. Keynote speech, 28th annual meeting of the DGfS, University of Bielefeld.

Ide, Nancy, 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)*, 825–830, Athens.

Lehmberg, Timm, Christian Chiarcos, Erhard Hinrichs, Georg Rehm, and Andreas Witt (2007a). Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System. In Sara Schmidt, Ray Siemens, Amit Kumar, and John Unsworth (eds.), *Digital Humanities 2007*, 164–166, ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Lehmberg, Timm, Christian Chiarcos, Georg Rehm, and Andreas Witt (2007b). Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer (eds.), *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, 93–102, Tübingen: Narr.

Lehmberg, Timm, Georg Rehm, Andreas Witt, and Felix Zimmermann (2008). Preserving Linguistic Resources: Licensing – Privacy Issues – Mashups. *Library Trends* In print.

Lehmberg, Timm and Kai Wörner (2007). Annotation Standards. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics*, Handbücher zur Sprach- und Kommunikationswissenschaft (HSK), Berlin, New York: de Gruyter, in press.

Lezius, Wolfgang (2002). *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Ph.D. thesis, University of Stuttgart.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):313–330.

Rehm, Georg, Richard Eckart, and Christian Chiarcos (2007a). An OWL- and XQuery-Based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nicolai Nikolov (eds.), *International Conference Recent Advances in Natural Language Processing (RANLP 2007)*, 510–514, Borovets, Bulgaria.

Rehm, Georg, Richard Eckart, Christian Chiarcos, and Johannes Dellert (2008a). Ontology-Based XQuery'ing of XML-Encoded Language Resources on Multiple Annotation Layers. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Rehm, Georg, Oliver Schonefeld, Andreas Witt, Timm Lehmberg, Christian Chiarcos, Hanan Bechara, Florian Eishold, Kilian Evang, Magdalena Leshtanska, Aleksandar Savkov, and Matthias Stark (2008b). The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Rehm, Georg, Andreas Witt, and Lothar Lemnitzer (eds.) (2007b). *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*. Tübingen: Narr.

Rehm, Georg, Andreas Witt, Heike Zinsmeister, and Johannes Dellert (2007c). Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections. In Sara Schmidt, Ray Siemens, Amit Kumar, and John Unsworth (eds.), *Digital Humanities 2007*, 166–170, ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Rehm, Georg, Andreas Witt, Heike Zinsmeister, and Johannes Dellert (2007d). Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, number 1 in Northern European Association for Language Technology Proceedings Series, 127–138, Bergen, Norway.

Sampson, Geoffrey (1995). *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon.

Schiller, Arne, Simone Teufel, and Christine Thielen (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, University of Stuttgart, University of Tübingen.

Schmidt, Thomas (2005). Time Based Data Models and the Text Encoding Initiative's Guidelines for Transcription of Speech. *Working Papers in Multilingualism, Series B* 62.

Schmidt, Thomas, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs (2006). Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*, East Lansing, Michigan.

Soehn, Jan-Philipp, Heike Zinsmeister, and Georg Rehm (2008). Requirements of a User-Friendly, General-Purpose Corpus Query Interface. In Lou Burnard, Khalid Choukri, Georg Rehm, Thomas Schmidt, and Andreas Witt (eds.), *Proceedings of the LREC 2008 Workshop Sustainability of Language Resources and Tools for Natural Language Processing*, Marrakech, Morocco.

Sperberg-McQueen, C. M. and Lou Burnard (eds.) (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen.

Telljohann, Heike, Erhard Hinrichs, and Sandra Kübler (2004). The TüBa-D/Z Treebank – Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Trilsbeek, Paul and Peter Wittenburg (2006). Archiving Challenges. In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel (eds.), *Essentials of Language Documentation*, 311–335, Berlin, New York: Mouton de Gruyter.

Trippel, Thorsten (2004). Metadata for Time Aligned Corpora. In *Proceedings of the LREC Workshop: A Registry of Linguistic Data Categories within an Integrated Language Repository Area*, Lisbon.

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

Witt, Andreas, Oliver Schonefeld, Georg Rehm, Jonathan Khoo, and Kilian Evang (2007). On the Lossless Transformation of Single-File, Multi-Layer Annotations into Multi-Rooted Trees. In B. Tommie Usdin (ed.), *Proceedings of Extreme Markup Languages 2007*, Montréal, Canada.

Wörner, Kai, Andreas Witt, Georg Rehm, and Stefanie Dipper (2006). Modelling Linguistic Data Structures. In B. Tommie Usdin (ed.), *Proceedings of Extreme Markup Languages 2006*, Montréal, Canada.

Zimmermann, Felix and Timm Lehmberg (2007). Language Corpora – Copyright – Data Protection: The Legal Point of View. In Sara Schmidt, Ray Siemens, Amit Kumar, and John Unsworth (eds.), *Digital Humanities 2007*, 162–164, ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.