Procesamiento del lenguaje Natural, nº 41 (2008), pp. 155-162 recibido 7-05-2008; aceptado 16-06-2008

A Web-Platform for Preserving, Exploring, Visualising, and Querying Linguistic Corpora and other Resources

Plataforma web para el mantenimiento, exploración, visulización y búsqueda en corpus lingüísticos y en otros recursos

Georg Rehm,
Oliver Schonefeld, Andreas Witt
SFB 441 Linguistic Data Structures
University of Tübingen
Nauklerstrasse 35
72074 Tübingen, Germany
georg.rehm@uni-tuebingen.de
oliver.schonefeld@uni-tuebingen.de
andreas.witt@uni-tuebingen.de

Christian Chiarcos
SFB 632 Information Structure
University of Potsdam
Karl-Liebknecht-Strasse 24–25
14476 Potsdam, Germany
chiarcos@uni-potsdam.de

Timm Lehmberg
SFB 538 Multilingualism
University of Hamburg
Max-Brauer-Allee 60
22765 Hamburg, Germany
timm.lehmberg@uni-hamburg.de

Resumen: Presentamos SPLICR, una plataforma de sostenibilidad para corpus y recursos lingüísticos basada en web. El sistema está destinado a personas que trabajan en el campo de la lingüística o de la lingüística computacional. Consiste en una base de datos extensa para metadatos que puede ser explorada para buscar recursos lingüísticos, que pudieran ser apropiados para las necesidades específicas de una investigación. SPLICR también ofrece una interfaz gráfica, que permite a los usuarios buscar y visualizar los corpus. El proyecto, en el que se ha desarollado el sistema, aspira a archivar de modo sostenible aproximadamente sesenta recursos lingüísticos, que han sido construidos mediante la colaboración de tres centros de investigación. Nuestro proyecto tiene dos metas principales: (a) Procesar y archivar recursos de forma sostenible, de manera que los recursos sigan siendo accesibles para la comunidad científica dentro de cinco, diez, o incluso veinte años. (b) El permitir a los investigadores buscar en los recursos tanto a nivel de metadatos como a nivel de anotaciones lingüísticas. En términos más generales, nuestro objetivo es proporcionar soluciones que posibiliten la interoperabilidad, reutilización y sostenibilidad de compilaciones heterogéneas de recursos de lenguaje.
Palabras clave: sostenibilidad, mantenimiento, búsqueda, corpus, recursos, XML

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 a graphical interface that enables users to query and to visualise corpora. The project in which the system is 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.
Keywords: Sustainability, Preservation, Querying, Corpora, Resources, XML

ISSN: 1135-5948

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 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. 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 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 front-end (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; Sperberg-McQueen and Burnard, 2002; Wörner et al., 2006; Lehmberg and Wörner, 2007) 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.

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 (Witt et al., 2007; Rehm et al., 2008a); 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 (Rehm et al., 2007a; Rehm et al., 2008b).

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 (Zimmermann and Lehmberg, 2007; Lehmberg et al., 2007a; Lehmberg et al., 2007b; Lehmberg et al., 2008; Rehm et al., 2007b): 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., 2007b; Rehm et al., 2007c).

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 (Trippel, 2004; Himmelmann, 2006). One eTEI file contains information on one of the following levels: (1) *setting* (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*; (5) *a corpus* (consists of primary data with one or more annotation levels).

We devised a workflow that helps users edit eTEI records (Rehm et al., 2008a). The workflow's primary components are the eTEI DTD and the Oxygen XML editor. Based on structured 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 technology. 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.

[1] In the file vault area of the system, 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.

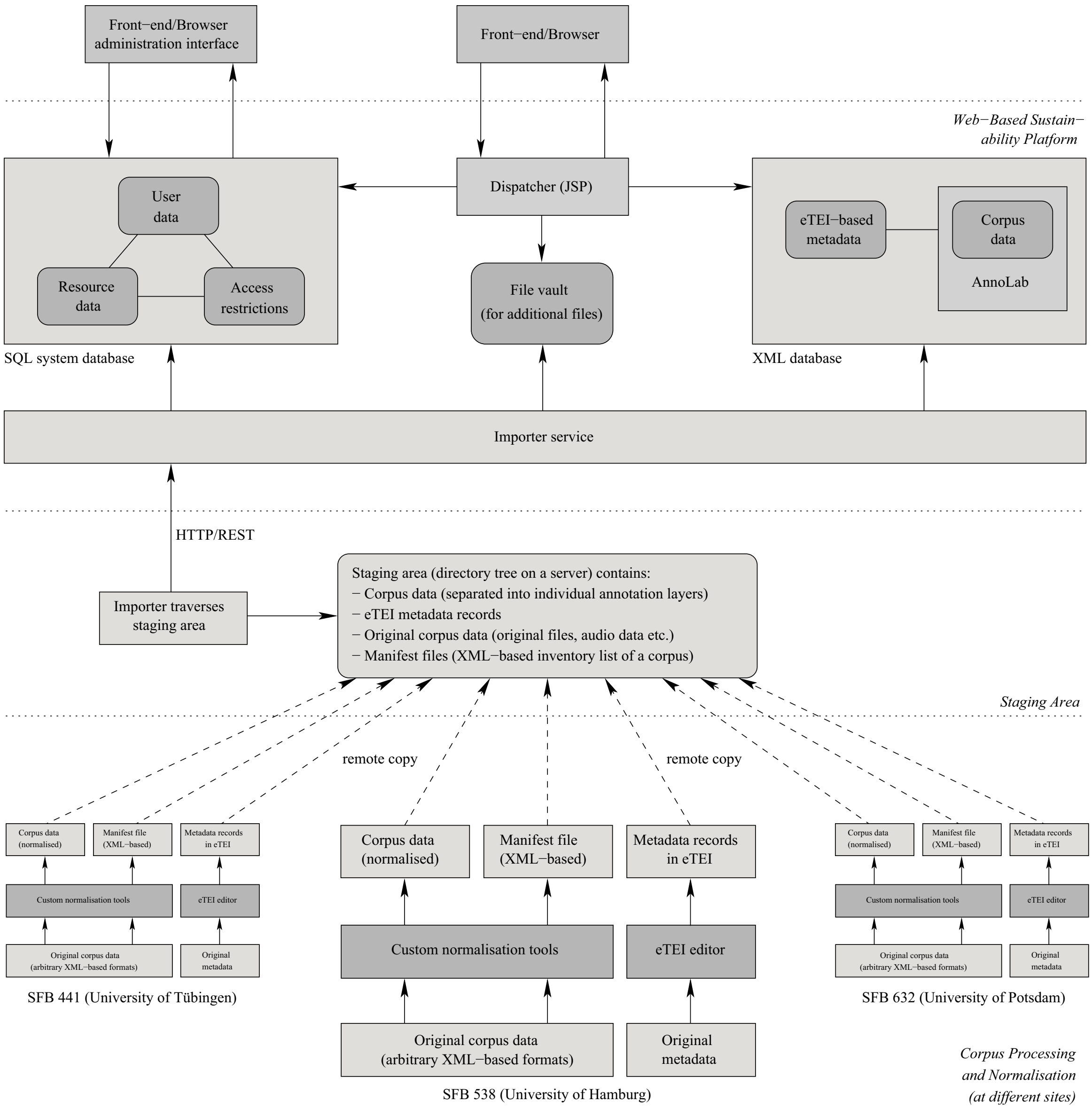

Front–end/Browser administration interface
Front–end/Browser
Web–Based Sustain–ability Platform
User data
Resource data
Access restrictions
SQL system database
Dispatcher (JSP)
File vault (for additional files)
eTEI–based metadata
Corpus data
AnnoLab
XML database
Importer service
HTTP/REST
Importer traverses 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 editor
Original corpus data (arbitrary XML–based formats)
Original metadata
SFB 441 (University of Tübingen)
Corpus data (normalised)
Manifest file (XML–based)
Metadata records in eTEI
Custom normalisation tools
eTEI editor
Original corpus data (arbitrary XML–based formats)
Original metadata
SFB 538 (University of Hamburg)
Corpus data (normalised)
Manifest file (XML–based)
Metadata records in eTEI
Custom normalisation tools
eTEI editor
Original corpus data (arbitrary XML–based formats)
Original metadata
SFB 632 (University of Potsdam)
Corpus Processing and Normalisation (at different sites)

Figure 1: Resource normalisation and SPLICR's staging area

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 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.

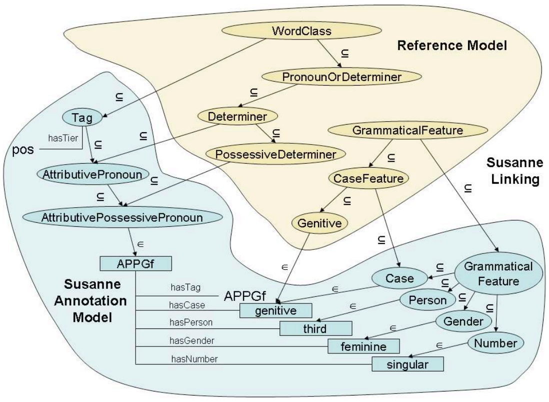


Figure 2: The Susanne tag `APPGf`, its representation within the annotation model and linking with the reference model

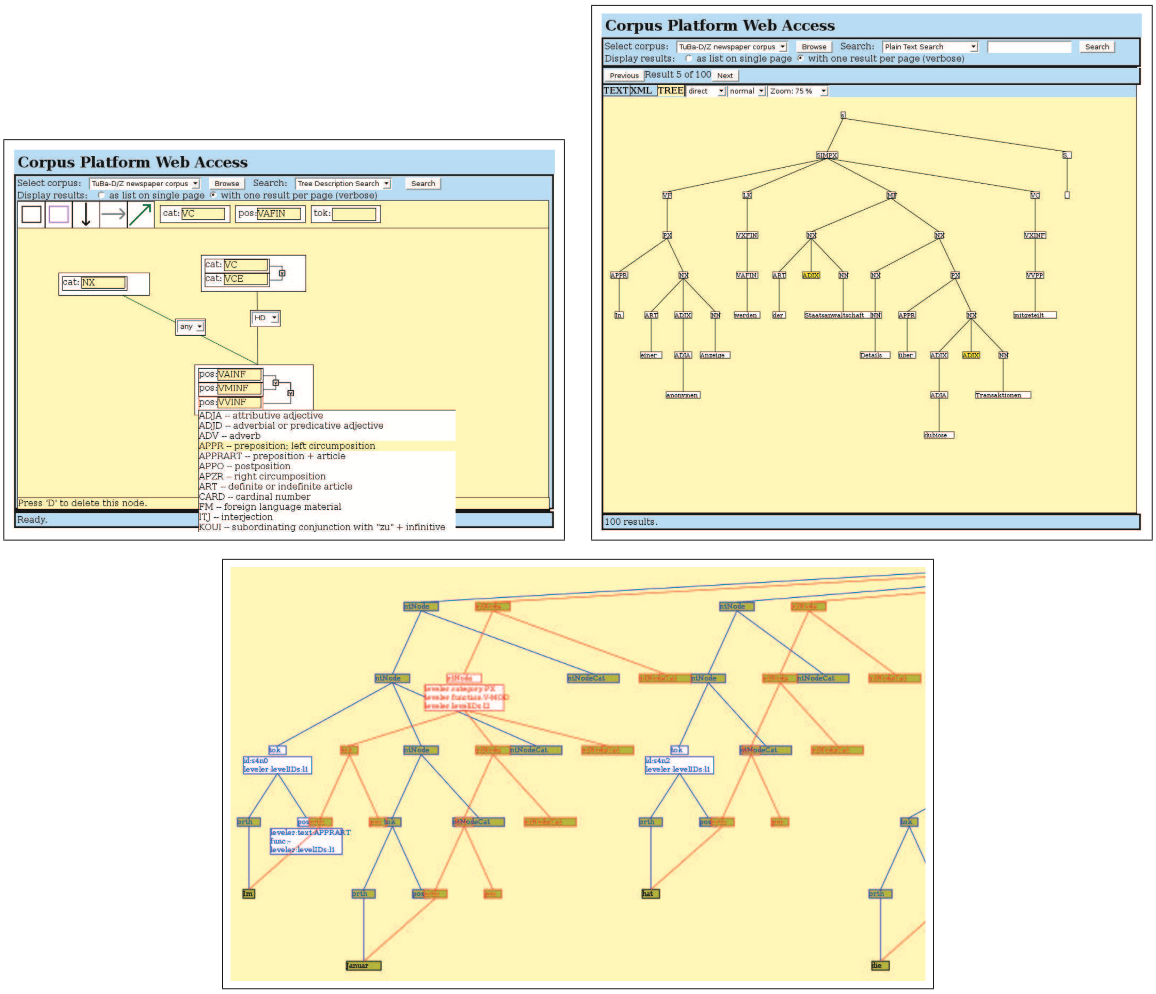
3.2. Ontologies of Linguistic Annotation

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 SFB632 annotation standard (Dipper et al., 2007). It covers reference specifications for word classes, and morpho-syntax (Chiarcos, 2007), and is currently extended to syntax and information structure. The 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 annotation model is an ontology that represents one specific annotation scheme (see figure 2). We built, among others, annotation models for the SFB632 annotation format (Dipper et al., 2007) used in typological research, TIGER/STTS (Schiller et al., 1999; Brants et al., 2003), 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 `Verb` or `Noun` 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.

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 spe-



Corpus Platform Web Access

Select corpus: TuBa-D/Z newspaper corpus Browse Search: Tree Description Search Search

Display results: ○ as list on single page ● with one result per page (verbose)

cat: VC pos: VAFIN tok:

cat: NX

cat: VC
cat: VCE

any

HD

pos: VAINF
pos: VMINF
pos: VVINF

ADJA -- attributive adjective
ADJD -- adverbial or predicative adjective
ADV -- adverb
APPR -- preposition; left circumposition
APPRART -- preposition + article
APPO -- postposition
APZR -- right circumposition
ART -- definite or indefinite article
CARD -- cardinal number
FM -- foreign language material
ITJ -- interjection
KOUI -- subordinating conjunction with "zu" + infinitive

Press 'D' to delete this node.

Ready.

Corpus Platform Web Access

Select corpus: TuBa-D/Z newspaper corpus Browse Search: Plain Text Search Search

Display results: ○ as list on single page ● with one result per page (verbose)

Previous Result 5 of 100 Next

TEXT XML TREE direct normal Zoom: 75 %

100 results.

Figure 3: The front-end in tree fragment query mode (upper left), single-tree corpus browsing mode (upper right) and multi-rooted tree display mode (below)

cific 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 (`http://www.prototypejs.org`) and script.aculo.us (`http://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 matches encoded in an XML format, which is then interpreted by a variety of display modules. Four different major display modes are already implemented: plain text view, XML view, graphical tree view and timeline view.

The tree fragment query editor (figure 3) 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., 2008b) 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.

Acknowledgments

The authors would like to thank Solange Otermin and Wolfgang Maier for translating this paper's abstract into Spanish. We would also like to thank the three anonymous reviewers for their helpful comments.

References

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

S. Brants, S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit. 2003. TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation.*

J. Carletta, J. Kilgour, T. J. O'Donnell, S. Evert, and H. Voormann. 2003. The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets. In *Proc. of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML).*

C. Chiarcos. 2007. An Ontology of Linguistic Annotation: Word Classes and Morphology. In *Proc. of DIALOG 2007.*

S. Dipper, E. Hinrichs, T. Schmidt, A. Wagner, and A. Witt. 2006. Sustainability of Linguistic Resources. In E. Hinrichs, N. Ide, M. Palmer, and J. Pustejovsky, editors, *Proc. of the LREC 2006 Workshop Merging and Layering Linguistic Information*, pages 48–54, Genoa, Italy, May.

S. Dipper, M. Götze, and S. Skopeteas, editors. 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, volume 7 of *ISIS*.

R. Eckart and E. Teich. 2007. An XML-Based Data Model for Flexible Representation and Query of Linguistically Interpreted Corpora. In G. Rehm, A. Witt, and L. Lemnitzer, editors, *Data Structures for Linguistic Resources and Applications.* Gunter Narr, Tübingen, Germany.

S. Farrar and D. T. Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *GLOT International*, 3:97–100.

N. P. Himmelmann. 2006. Daten und Datenhuberei. Keynote speech, 28th annual meeting of the DGfS, University of Bielefeld, February.

N. Ide, P. Bonhomme, and L. Romary. 2000. XCES: An XML-based Standard for Linguistic Corpora. In *Proc. of the Second Language Resources and Evaluation Conf. (LREC)*, pages 825–830, Athens.

T. Lehmberg and K. Wörner. 2007. Annotation Standards. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics*, Handbücher zur Sprach- und Kommunikationswissenschaft (HSK). de Gruyter, Berlin, New York. In press.

T. Lehmberg, C. Chiarcos, E. Hinrichs, G. Rehm, and A. Witt. 2007a. Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System. In *Digital Humanities 2007*, pages 164–166, Urbana-Champaign, IL, USA, June. ACH, ALLC.

T. Lehmberg, C. Chiarcos, G. Rehm, and A. Witt. 2007b. Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In G. Rehm, A. Witt, and L. Lemnitzer, editors, *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proc. of the Biennial GLDV Conf. 2007*, pages 93–102. Gunter Narr, Tübingen.

T. Lehmberg, G. Rehm, A. Witt, and F. Zimmermann. 2008. Preserving Linguistic Resources: Licensing – Privacy Issues – Mashups. *Library Trends*. In print.

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

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June.

G. Rehm, R. Eckart, and C. Chiarcos. 2007a. An OWL- and XQuery-Based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora. In *Int. Conf. Recent Advances in Natural Language Processing (RANLP 2007)*, pages 510–514, Borovets, Bulgaria, September.

G. Rehm, A. Witt, H. Zinsmeister, and J. Dellert. 2007b. Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections. In *Digital Humanities 2007*, pages 166–170, Urbana-Champaign, IL, USA, June. ACH, ALLC.

G. Rehm, A. Witt, H. Zinsmeister, and J. Dellert. 2007c. Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications. In *Proc. of the Sixth Int. Workshop on Treebanks and Linguistic Theories (TLT 2007)*, number 1 in Northern European Association for Language Technology Proc. Series, pages 127–138, Bergen, Norway, December.

G. Rehm, O. Schonefeld, A. Witt, T. Lehmberg, C. Chiarcos, H. Bechara, F. Eishold, K. Evang, M. Leshtanska, Aleksandar Savkov, and Matthias Stark. 2008a. The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources. In *Proc. of the 6th Language Resources and Evaluation Conf. (LREC 2008)*, Marrakech, Morocco, May.

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

G. Sampson. 1995. *English for the Computer. The SUSANNE Corpus and Analytic Scheme.* Clarendon, Oxford.

A. Schiller, S. Teufel, and C. Stockert. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, University of Stuttgart, University of Tübingen.

T. Schmidt, C. Chiarcos, T. Lehmberg, G. Rehm, A. Witt, and E. Hinrichs. 2006. Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *Proc. of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*, East Lansing, Michigan, June.

T. Schmidt. 2005. Time Based Data Models and the Text Encoding Initiative's Guidelines for Transcription of Speech. *Working Papers in Multilingualism, Series B*, 62.

J.-P. Soehn, H. Zinsmeister, and G. Rehm. 2008. Requirements of a User-Friendly, General-Purpose Corpus Query Interface. In L. Burnard, K. Choukri, G. Rehm, T. Schmidt, and A. Witt, editors, *Proc. of the LREC 2008 Workshop Sustainability of Language Resources and Tools for Natural Language Processing*, Marrakech, Morocco, May 31.

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

P. Trilsbeek and P. Wittenburg. 2006. Archiving Challenges. In J. Gippert, N. P. Himmelmann, and U. Mosel, editors, *Essentials of Language Documentation*, pages 311–335. Mouton de Gruyter, Berlin, New York.

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

A. Witt, O. Schonefeld, G. Rehm, J. Khoo, and K. Evang. 2007. On the Lossless Transformation of Single-File, Multi-Layer Annotations into Multi-Rooted Trees. In B. T. Usdin, editor, *Proc. of Extreme Markup Languages 2007*, Montréal, Canada, August.

K. Wörner, A. Witt, G. Rehm, and S. Dipper. 2006. Modelling Linguistic Data Structures. In B. T. Usdin, editor, *Proc. of Extreme Markup Languages 2006*, Montréal, Canada, August.

F. Zimmermann and T. Lehmberg. 2007. Language Corpora – Copyright – Data Protection: The Legal Point of View. In *Digital Humanities 2007*, pages 162–164, Urbana-Champaign, IL, USA, June. ACH, ALLC.