Published in: Proceedings of Extreme Markup Languages 2006, Montreal, Kanada. - Montreal, 2006.

Proceedings of Extreme Markup Languages®

Master Bibliography | Author Index | Topic Index | Date Index | Proceedings Home

Modelling Linguistic Data Structures

Kai Wörner
Andreas Witt
Georg Rehm
Stefanie Dipper

Abstract

Linguistic corpora have been annotated by means of SGML-based markup languages for almost 20 years. We can, very roughly, differentiate between three distinct evolutionary stages of markup technologies. (1) Originally, single SGML tree-based document instances were deemed sufficient for the representation of linguistic structures. (2) Linguists began to realize that alternatives and extensions to the traditional model are needed. Formalisms such as, for example, NITE were proposed: the NITE Object Model (NOM) consists of multi-rooted trees. (3) We are now on the threshold of the third evolutionary stage: even NITE's very flexible approach is not suited for all linguistic purposes. As some structures, such as these, cannot be modeled by multi-rooted trees, an even more flexible approach is needed in order to provide a generic annotation format that is able to represent genuinely arbitrary linguistic data structures.

Keywords: Trees/Graphs; Modeling; Markup Languages

Table of Contents

Kai Wörner

Kai Wörner has a master's degree in German linguistics from the Giessen University, Germany. After working as a freelance programmer and web-designer, he is now working at the collaborative research center on multilingualism (SFB 538) in Hamburg, dealing with computer-based methods for the collection and processing of multilingual data.

Andreas Witt

From 1996 to 2006, Andreas Witt has taught at Bielefeld University, Germany in the field of 'text technology'. His research interests include the combination of computational linguistics and markup technologies, schema languages, and corpus annotation. Since April 2006 Dr. Witt is engaged in a joint project of the Universities Hamburg, Potsdam/Berlin and Tübingen. on "Sustainability of linguistic data"

Georg Rehm

Georg Rehm works in Tübingen University's collaborative research centre Linguistic Data Structures in a project that develops the foundations for sustainable linguistic resources. He holds a PhD in Applied and Computational Linguistics and has been working with SGML and related technologies in the context of Natural Language Processing (especially with regard to text and corpus analysis as well as ontologies) since 1995.

Stefanie Dipper

Stefanie Dipper is a computational linguist at the University of Potsdam. She holds a Magister degree in linguistics, computer science and psychology from the University of Tübingen, Germany. From 1997 to 2002 she was a research assistant at IMS Stuttgart, where she was involved in the implementation of a large-scale German LFG grammar (Pargram project) and in the annotation of a German treebank, TIGER. She obtained the degree of Dr. phil. from the University of Stuttgart in 2003. She joined the Applied CL Group of Potsdam in Sep 2003, and became a member of the SFB 632 (project D1) and the project SUMMaR. Her research interests include German syntax, discourse structure, grammar engineering, and corpus linguistics.

XML Source PDF (for print) Author Package Typeset PDF

Modelling Linguistic Data Structures

Kai Wörner [Hamburg University]
Andreas Witt [Tübingen University]
Georg Rehm [Tübingen University]
Stefanie Dipper [Potsdam University]

Extreme Markup Languages 2006® (Montréal, Québec)

Introduction

Linguistic corpora have been annotated by means of SGML-based markup languages since the formalism was standardised in the 1980ies. We can, very roughly, differentiate between three distinct evolutionary stages of markup technologies. (a) Originally, single SGML document instances were deemed to be sufficient for the representation of linguistic structures. In the late 1980ies, early 1990ies, the Text Encoding Initiative proposed a complex approach that consists of multiple markup languages that can be added to SGML instances in a modular way in order to provide hundreds of elements that can be used for linguistic purposes. The TEI guidelines described two ways of annotating conceptually different textual data. These are the rather traditional approach of using hierarchical structures (all document grammars used in an instance are merged into a single DTD) and a method to treat the conceptually different TEI DTD modules as separate document grammars (concurrent markup). Because of technical restrictions only the first approach has been used in practice. Even the TEI's influential "Gentle Introduction to SGML"

mentions the central and very problematic aspect of concurrent markup: traditional SGML- and XML-documents provide trees only, but linguistic data contains overlapping structures on a regular basis – these structures cannot be modelled within XML's paradigm of nested element trees. (see [Sperberg-McQueen and Burnard 1994], [Barnard et al. (1995)], [Witt (2004)], and [DeRose (2004)]) It is true, though, that SGML's CONCUR feature can be used to enable the use of multiple markup languages in a single document, but CONCUR has, to our knowledge, never been fully implemented and failed to find its way into the XML standard. (see [Hilbert et al.(2005)] and [Schonefeld and Witt (2006)]) (b) In the second evolutionary stage, linguists began to realise that alternatives and extensions to the traditional model are needed in order to capture conceptually different structures in an adequate way (for example, syntax, intonation, morphology and syllable structure of a transcribed utterance). To overcome this problem two groups of formalisms have been proposed. The rather general graph-based approaches such as annotation graphs employ a directed graph structure that is able to link the arcs without any restrictions. Most applications of annotation graphs use a predominant arc, the timeline, to anchor all the link-points. The second group can be characterised as tree-based. For example, the NITE Object Model (NOM, [Carletta et al. 2003]) consists of multi-rooted trees, i.e., multiple element trees can refer to one set of primary data that is optionally augmented by a timeline. The multiple trees are represented in multiple files and can be viewed as stand-off annotation. (c) We have reviewed all major linguistic annotation frameworks with the aim of applying them to several dozen heterogenous, annotated linguistic corpora that have been created in the past ten years. It is our opinion that we are now on the threshold of the third evolutionary stage: even NITE's very flexible approach is not suited for all linguistic purposes. For example, it is perfectly reasonable for a linguist to annotate both an orthographic transcription and a phonetic representation of a spoken dialogue between two people. There are correspondences between these two layers, but, eventually, they are independent and have to be treated as primary data each. As structures such as these cannot be modelled by the NITE Object Model, an even more flexible approach is needed in order to provide a generic annotation format that is able to represent genuinely arbitrary linguistic data structures.

Goal of our project

The scenario sketched in the introduction is the starting point of the approach we describe in this contribution. One of the major goals of our project is to devise a generic data format that is able to consolidate conceptually different markup languages in order to act as a kind of least common denominator. The generic data format is still work in progress and will be tested with several dozen corpora, based on three different annotation frameworks (the timeline-based stand-off format *Exmaralda* ([Schmidt 2004]), the hierarchical format *Tusnelda* that is based on the TEI ([Sperberg-McQueen and Burnard 1994]), and *Paula* that resembles the Linguistic Annotation Framework ([Ide et al. 2003])) that will be transformed into the meta-format. An ontology of linguistic terms and concepts will be used to take care of the commonalities and differences of the three markup languages that have been used to annotate the source corpora. With the help of the generic data format and the ontology it will be possible to query, to compare, and to analyse heterogenous sets of corpora, or, rather, arbitrary sets of XML-annotated data, in a uniform way.

Structure of the paper

This paper begins with an introduction of the three annotation frameworks that are used at the research centres involved in the development of the generic data format. Afterwards our considerations concerning the generic meta/exchange format that should be able to subsume the three frameworks are presented, also mentioning unsolved problems and open questions.

The problem we are trying to tackle with devising the generic data format and a set of accompanying tools is a rather general one, as the need to consolidate large data sets marked up using heterogeneous annotation frameworks is not necessarily related to the discipline of linguistics exclusively. The generic data format will not be restricted to the area of linguistics only. We hope that it evolves to be an approach and a set of tools that is universally applicable.

Three approaches to linguistic markup

Exmaralda

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 1999]), 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 simple 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. Although the structure of the XML-files is given in a DTD, the graph model does not make use of XML's strength to formulate constraints on hierarchical relations and defining tag sets or annotation vocabularies.

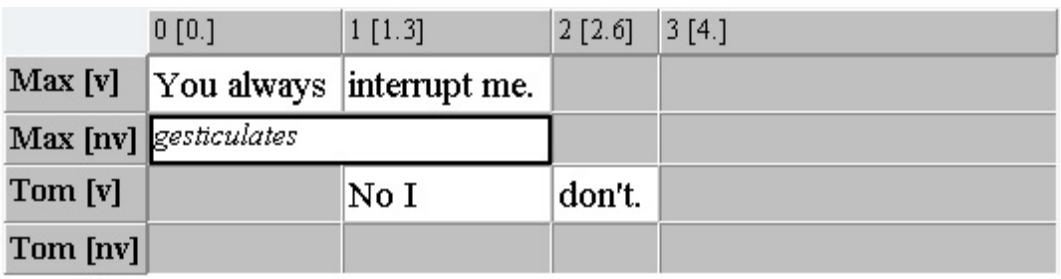
Conversion filters have been developed for legacy data. Due to a lack of documentation and several inconsistencies in these older corpora, however, a complete conversion cannot be accomplished automatically, but requires a substantial amount of manual post-editing.

At the present time, linguistic data represented in the EXMARaLDA data format is usually created with the help of the EXMARaLDA Partitur-Editor, a tier-based tool presenting the transcription to the user as a musical score supporting the creation of links between the transcription and the underlying digitized audio 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. The EXMARaLDA tools are described in detail in [Schmidt and Wörner 2005] and in various materials available from the project website (http://www.rrz.uni-hamburg.de/exmaralda).

The transfer from the directed graph structure of transcription-graphs in EXMARaLDA to a data model which is hierarchy-oriented (e.g., single-rooted or multi-rooted trees) has to be accomplished via the graph's ordered nodes that establish the structure and are the only valid markers as to how annotations are linked to textual content. These nodes are translated into anchor points in the "root"-XML-file. The segments of the textual content link to their start- and end anchors are maintained in a separate XML file. Annotations on the textual content again link to these segments via pointers, so that the relations between the text and the annotations do not have to be calculated by means of the anchors.

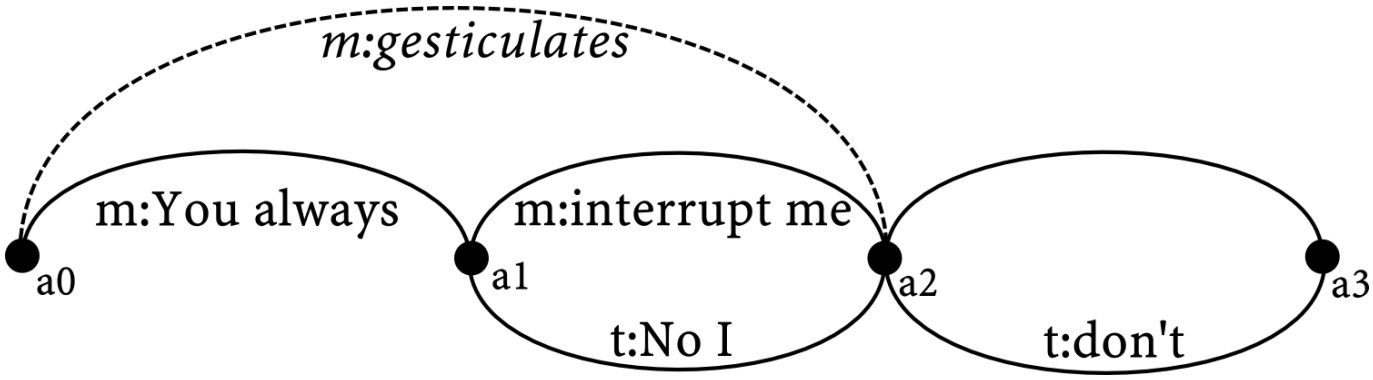
The following example shows a typical transcription example with two speakers, annotation and overlap and the according transcription graph in exmaralda:

Figure 1: Partiture



[Link to open this graphic in a separate page]

Figure 2: Graph



[Link to open this graphic in a separate page]

The inline XML-representation in EXMARaLDA's basic transcription format would look like this:

```
<common-timeline> <tli
            id="T0" time="0.0"/> <tli
            id="T1"
            time="1.3333333333333333"/> <tli
            id="T2"
            time="2.666666666666667"/> <tli
            id="T3" time="4.0"/>
            </common-timeline> <tier id="TIE0"
            speaker="SPK0" category="v"
            type="t" display-name="Max [v]">
            <event start="T0"
            end="T1">You always </event>
            <event start="T1"
            end="T2">interrupt me. </event>
            </tier>
```

Following is the XML file representing the anchors for this example:

```
<anchors> <anchor
            id="a0"/> <anchor
            id="a1"/> <anchor
            id="a2"/> <anchor
            id="a3"/> </anchors>
```

The information about “time” is stripped in the conversion from a timeline to a list of anchors and would have to be hold in a separate file. Connections between layers that represent the utterances of multiple speakers can only be calculated through these anchors.

Each textual segment has its own id and links to an anchor marking the beginning and one marking the end.

```
<text
            spk="max"> <seg id="tm0"
            href="anchors.xml#a0..#a1"> You always
            </seg> <seg id="tm1"
            href="anchors.xml#a1..#a2"> interrupt me
            </seg> </text>
```

The annotation-layers are held in separate files, linking to the ids of the textual segments, not to the anchors, thus linking the annotation to unique speakers.

```
<annotation type="nv"
            spk="max"> <seg
            href="spk_max.xml#tm0..#tm1">
            gesticulates</seg> </annotation>
```

An annotation not linked to a certain speaker would be linked to the anchors directly.

The transcription-graph model of EXMARaLDA does not provide any means of adding hierarchical annotations to transcriptions. Annotations can only be applied to segments between the nodes of the graph and belong either to a speaker (and thus to language data) or refer to things apart from the actual speech. The textual layers are open to further annotation, though. As long as transcription-graphs do not allow for hierarchical annotation, applying such annotation would result to information loss when converting back from the meta-format. EXMARaLDAs extended format, that segments transcriptions further according to the transcription rules used, could be converted to the meta-format, too, since it follows the same model as the basic format. Conversion from this format would result in XML-files for each segmentation rule (like utterances, words, phonemes), so that annotations could be applied to any segment. Converting these annotations back to EXMARaLDA is not possible since the format does not account for this type of annotation yet and there is also no software tool to display segmented transcriptions yet.

Tusnelda

Tusnelda is an acronym for the German translation of "Tübingen collection of reusable, empirical, linguistic data structures". This collection contains heterogeneous corpora that differ with respect to several aspects (e.g., annotated languages, text types, kind of annotated, language-related information). Nonetheless a common annotation scheme, also called Tusnelda, has been developed several years ago. The development of the Tusnelda annotation scheme was heavily influenced by the work of the Text Encoding Initiative (TEI) and by the TEI-influenced Corpus Encoding Standard (XCES).

In contrast to the Exmaralda data format, Tusnelda does make use of a hierarchical data model, and all the Tusnelda corpora consist of XML-files which have been validated against the Tusnelda Document Type Definition.

The following example shows a Tusnelda file. The linguistic aspects of this extract of the Tibetean corpus can be found in **[Wagner and Zeisler (2004)]**.

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

This extract shows a standard XML-structure. However, a closer look reveals implicit information. The natural (and intended) way of interpreting the transcripted and annotated utterance is to relate the node <pos> to the node <orth>, i.e., to relate a transcribed word with information on its part of speech. From an XML-oriented point of view, however, the nodes <pos> and <orth> are simply adjacent nodes. Another example of adjacent nodes are the first and the second token <tok>. Hence, in the case of <tok> two neighboring tags represent a sequence but in other cases (e.g., <orth>and <pos>, or a sequence of <realComplement>) two adjacent tags provide different additional information with regard to the same text.

The general data format should avoid ambiguities of this kind. Of course, a general format without these

ambiguities would lead to the necessity of transforming the Tusnelda corpora into the new format. Ideally, this transformation should be able to resolve the described ambiguities automatically.

Paula

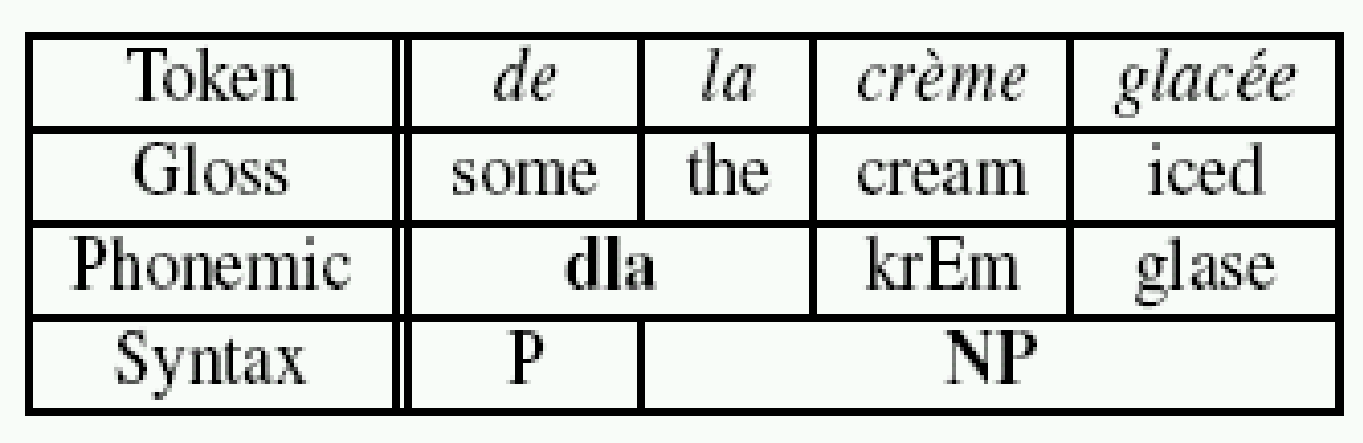
The interchange format PAULA has been developed for empirical, data-based research on information structure, a linguistic phenomenon that involves various linguistic levels, such as syntax, phonology, semantics. As a consequence, the data which serve as the basis of this research are marked up with different kinds of annotations: syntax trees or graphs, segment-based phonological properties, etc. The annotations are created by means of different, task-specific annotation tools: EXMARaLDA for segment-based annotations[1], [1] annotate for syntax graphs[2], [2], MMAX for anaphoric relations[3], [3], and RST Tool for discourse-relational trees[4], [4].

For instance, the newspaper articles in the Potsdam Commentary Corpus (PCC) have been annotated on multiple layers: morphology, part of speech, syntax, anaphoric relations, discourse relations, and information-structural properties. To allow for searching for correlations or interactions between the different annotation layers, the output representations of the task-specific annotation tools are converted to the flexible representation format PAULA that is inspired by the Linguistic Annotation Framework (**[Ide and Romary (2001)]**) and defines abstract XML elements and attributes that can represent different annotation types, such as graphs, pointers, and tier segments.

PAULA currently feeds three applications. First, for manual inspection of the data at multiple levels, data and its annotations are imported into the linguistic database ANNIS, which provides viewing and searching facilities (**[Dipper et al. 2004]**). Second, for annotation mining, i.e., automatic computation of feature correlations, PAULA representations are converted to the Weka .arff file format; Weka is a collection of machine learning algorithms for data mining tasks (**[Frank et al. (2005)]**). Third, we use PAULA to represent linguistic and statistical data in a text summarization system (**[Stede et al. (To appear)]**).

In the context of information-structural research, segments or graphs that annotations are attached to overlap quite often. The following example features an overlap between the phonemic and syntactic levels: at the phonemic level (third tier), tokens 1 and 2, “de la” (‘of the’), are treated as one unit, whereas at the syntactic level, tokens 2-3, “la crème glacée” (‘the ice-cream’), form an NP constituent (fourth tier)

Figure 3: Overlapping segments



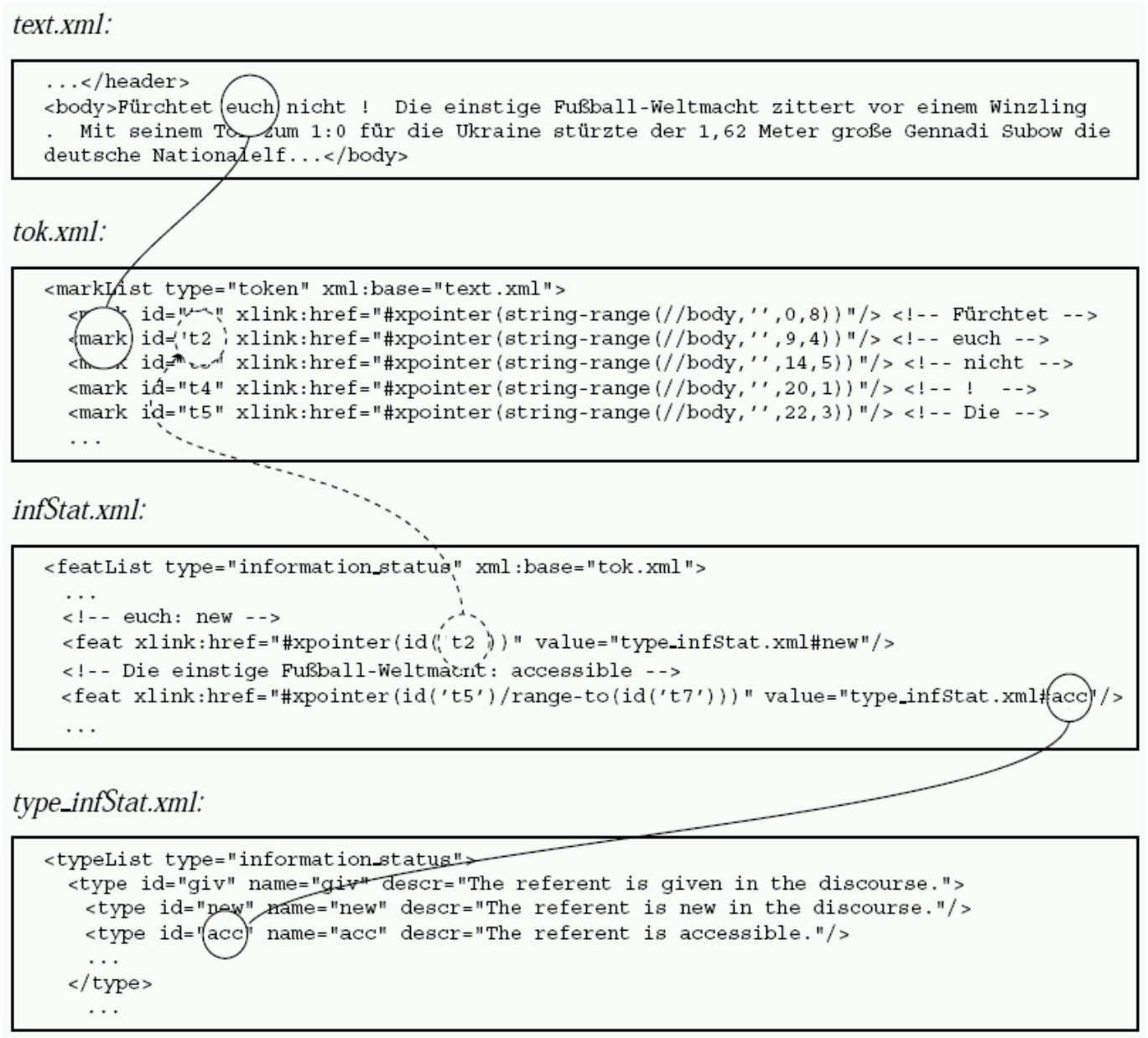
[Link to open this graphic in a separate page]

To account for such overlapping segments and for the heterogeneity of the data in general, PAULA uses an XML-based standoff architecture, each annotation type is stored in a separate file. Annotations refer to the source text or to other annotations, by means of XLinks and XPointers. We distinguish three different types of annotations: markables, structures, and features.

(i) Markables: <mark> tags specify text positions or spans of text (or spans of other markables) that can be annotated by linguistic information. For instance, <mark> tags might indicate tokens by specifying ranges of the source text, cf. figure 4.

(ii) Features: <feat> tags specify information annotated to markables, which are referred to by xlink attributes. The type of information (e.g., "information status") is encoded by an attribute "type". We adopt the idea of [Carletta et al. 2003] by assuming that admissible feature values (such as "new", marking information new in the discourse, or "NP", noun phrase) may be complex types and are organized in a type hierarchy, cf. figure 4 (the figure displays a simplified version of a type hierarchy, which actually has the form of a directed graph).

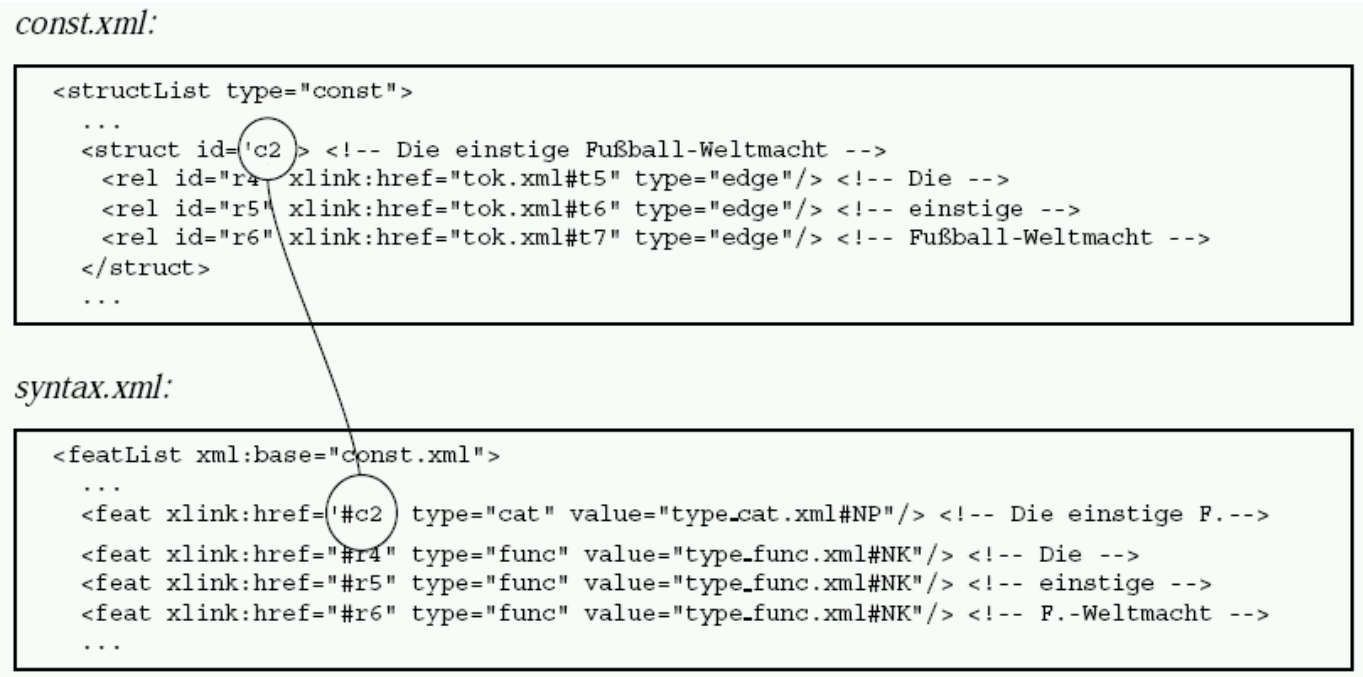
Figure 4: Markables, feature annotations, and type hierarchy



[Link to open this graphic in a separate page]

(iii) Structures: <struct> tags are special types of markables. Similar to <mark> tags, they specify objects that then can serve as anchors for annotations. A <struct> tag represents a complex anchor involving relations between arbitrarily many markables. For instance, the <struct> element in figure 5 defines a local subtree consisting of a mother node (<struct id="c2">) and three daughters (= the markables referenced by the xlink attributes).

Figure 5: Graph representation



[Link to open this graphic in a separate page]

PAULA furthermore encodes meta information, such as the object language, the speaker's name etc. Finally, PAULA allows the user to group related annotation levels; for instance, competing analyses can be stored as separate annotations and marked as alternatives.

The generic data format

We are currently in the process of devising a generic data format that will be able to subsume the properties of the three linguistic markup languages or frameworks mentioned in the previous sections. From the markup community's point of view, both the domain (linguistic data) as well as the three languages are regarded as case studies only; the resulting generic data format and the set of tools to be developed are meant to be a universal solution for the ubiquitous problem of finding a common format that is able to unify the specifics of conceptually different markup languages for the purpose of providing a sustainable archive of inherently diverse resources (see also **[Dipper et al. 2006]**, **[Schmidt et al. 2006]**).

There are several requirements the generic data format needs to meet, the most important of which is the ability to represent multi-rooted trees so that multiple annotation layers can be adequately modeled. The second requirement concerns non-redundancy: we would like to capture one or more potentially independent layers of primary data that can be augmented by additional layers that can, in turn, be interconnected in a flexible way using pointers. The generic data format needs to be able to model trees as well as and graphs, and it must be possible to anchor one or more annotation layers to a timeline. As the generic data format is to be based on XML, we plan to follow the approach of representing every single layer of annotation in a document instance of its own, cross-referenced with the primary data layer or layers. Metadata about the corpus itself as well as the document instances associated with a corpus will be maintained in an additional XML document instance (this approach is based on NITE-XML, see **[Carletta et al. 2003]**).

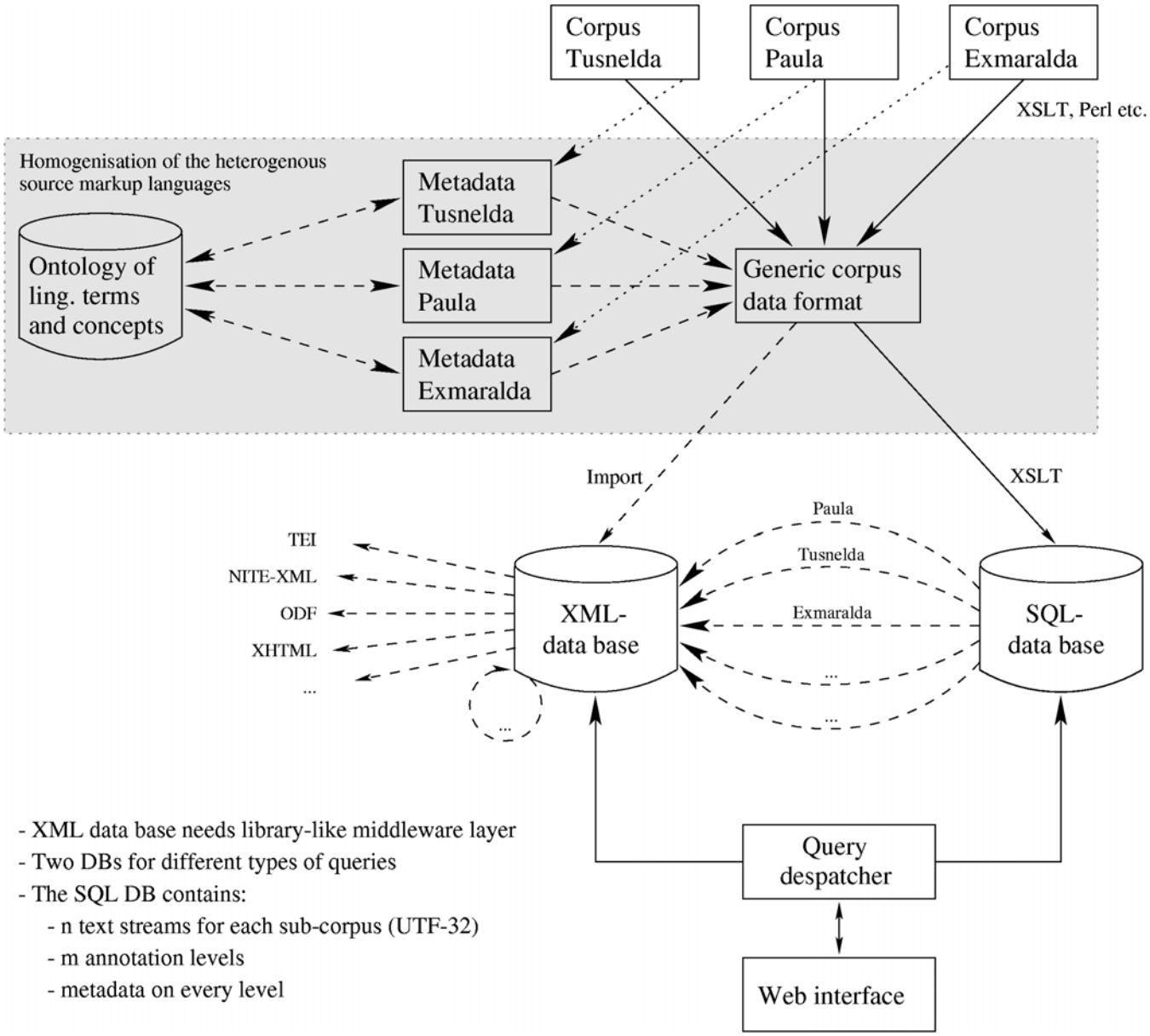
We plan to employ an ontology of linguistic terms and concepts (based on, for example, GOLD, see **[Farrar & Langendoen 2003]**) in order to map the tag and attribute names used in the source markup languages (SML) onto a class hierarchy of standardised concepts (ISO 24611): as all the source markup languages describe one particular domain of interest (in our case, linguistic corpora), it is crucial to relate the tag and attribute names to one another, so that, for example, unified search queries can be directed against the set

of corpora with the help of the ontology. If SML-1 uses the part-of-speech tag <N>, SML-2 uses <NOUN> and SML-3 uses <NOMEN>, we can map these different names onto the <noun> concept (subclass of the "part-of-speech" concept) and can use this ontological knowledge base in the search engine for the purpose of query expansion (see figure 6). Next to the rather general linguistic concepts ("utterance", "sentence", "word", "phrase", "noun", "verb", "discourse entity" etc.) the ontology will have to contain very specific concepts that are driven by the individual research questions examined in the projects the source data sets originate in. The ontology needs to be designed using an upper model comprising general concepts that is augmented by specific ontologies that are related to the domain of interest only.

Furthermore, developing XSLT stylesheets in order to transform a corpus annotated by an additional markup language into the generic data format should be as easy and straight-forward a task as possible (for example, by providing a pre-packaged XSLT library that simplifies the construction of the metadata file). The same holds for the process of transforming corpora already annotated in the generic data format into other formats. We plan to provide stylesheets that are able to export at least XHTML, TEI, DocBook, ODF, and PDF.

There are still several open questions: in a way, the approach sketched in this paper extendes the original scope of the XML standard by proposing to represent multiple layers of annotation for one or more sets of primary data in multiple files. We need to develop tools and libraries in order to enable us to work with large amounts of data annotated using this approach. Of utmost importance is the question whether it is possible to employ a native XML database (or a relational database) for the web-based query interface. Though a database would definitely provide fast response times, the question remains if it will be possible to modify or to serialise our generic data format, so that it is compatible with the established XML query formalisms. As an alternative, we would have to build an additional layer on top of the database engine that abstracts from the generic data format by means of an encapsulated process logic. A second problematic area regards the question how to interface the ontology of linguistic terms and concepts with the sets of metadata used in the individual source corpora. What if two or more corpora contain data annotated in markup languages that are, from a theoretical linguistics point-of-view, incompatible with each other (for example, if they are based on incompatible theoretical frameworks) – will it be possible to represent terms and concepts in the ontology that contradict each other?

Figure 6: Architecture



[Link to open this graphic in a separate page]

Summary

In the project "Sustainability of Linguistic Resoures", we develop a data format that will be able to subsume all current and future linguistic data, annotation and metadata of the research centres involved and possibly other researchers.

The data format will be able to represent multi-rooted trees, capture one or more layers of primary language data and additional layers of annotation. The layers will be linked with one another via pointers. Annotation data will be mapped onto a class hierarchy of linguistic concepts by an ontology to faciliate unified searches over the data.

There are still open questions related to the ontology as well as to storage, querying and the tools that need to be developed to facilitate working with the data.

Notes

1. http://www.rrz.uni-hamburg.de/exmaralda/
2. http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/

3. http://mmax.eml-research.de/
4. http://www.wagsoft.com/RSTTool/

Bibliography

[Barnard et al. (1995)] David Barnard, Lou Burnard, Jean-Pierre Gaspart, Lynne A. Price, C.M. Sperberg-McQueen, Giovanni Battista Varile. "Hierarchical Encoding of Text: Technical Problems and SGML Solutions." The Text Encoding Initiative: Background and Contents, Guest Editors Nancy Ide and Jean Vèronis = Computers and the Humanities 29/3 (1995) 211-231.

[Bird & Liberman 1999] Steven Bird & Marc Liberman *Annotation graphs as a framework for multidimensional linguistic data analysis.* in: Towards Standards and Tools for Discourse Tagging, Proceedings of the Workshop. Association for Computational Linguistics, 1999.

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

[DeRose (2004)] DeRose, Steven: *Markup Overlap: A Review and a Horse*, Extreme Markup Languages 2004 Conference Proceedings, Montreal, 2004

[Dipper et al. 2004] Stefanie Dipper, Michael Götze, Manfred Stede, and Tillmann Wegst. (2004) "ANNIS: A Linguistic Database for Exploring Information Structure". Interdisciplinary Studies on Information Structure (ISIS) (2004), 245-279.

[Dipper et al. 2006] Stefanie Dipper, Erhard Hinrichs, Thomas Schmidt, Andreas Wagner, 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 on "Merging and Layering Linguistic Information", Genoa 2006.

[Farrar & Langendoen 2003] Scott Farrar& D. Terry Langendoen. (2003) A linguistic ontology for the Semantic Web, GLOT International 7(3), 97-100.

[Frank et al. (2005)] Eibe Frank, Mark A. Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H. Witten, and Leonhard Trigg. "WEKA - A Machine Learning Workbench for Data Mining". In Oded Maimon and Lior Rokach, editors, The Data Mining and Knowledge Discovery Handbook (2005), 1305-1314.

[Hilbert et al.(2005)] Mirco Hilbert and Oliver Schonefeld and Andreas Witt: *Making CONCUR work*, Extreme Markup Languages 2005 Conference Proceedings, Montreal, 2005

[Ide and Romary (2001)] Nancy Ide and Laurent Romary. "Standards for Language Resources". Proceedings of the IRCS Workshop on Linguistic Database (2001), 141-149.

[Ide et al. 2003] Nancy Ide, Laurent Romary, and Eric de la Clergerie *International Standard for a Linguistic Annotation Framework.* Ide, N., Romary, L., de la Clergerie, E. In: Proc. HLT-NAACL'03 Workshop on the Software Engineering and Architecture of Language Technology, 2003.

[Schmidt 2004] Thomas Schmidt *EXMARaLDA - ein System zur computergestützten Diskurstranskription.*Schmidt, T. In: Mehler, Alexander / Lobin, Henning: Automatische Textanalyse, 2003.

[Schmidt and Wörner 2005] Thomas Schmidt, Kai Wörner. *Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA*. Gesprächsforschung Online 2005.

[Schmidt et al. 2006] Thomas Schmidt, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. *Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources.* In Proceedings of the E-MELD workshop 2006, June, 22 2006 Ypsilanti.

[Schonefeld and Witt (2006)] Oliver Schonefeld and Andreas Witt. *Towards validation of concurrent markup*, Extreme Markup Languages 2006 Conference Proceedings, Montreal, 2006

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

[Stede et al. (To appear)] Manfred Stede, Heike Bieler, Stefanie Dipper, and Arthit Suriyawongkul. "SUMMaR: Combining Linguistics and Statistics for Text Summarization". In Proceedings of the 17th European Conference on Artificial Intelligence.

[Wagner and Zeisler (2004)] Andreas Wagner and Bettina Zeisler. A syntactically annotated corpus of Tibetan. In: Proc. of LREC 2004, p. 1141–1144, Lisboa.

[Witt (2004)] Andreas Witt. Multiple hierarchies: new aspects of an old solution. In: Proceedings of Extreme Markup Languages. Montreal, 2004

Modelling Linguistic Data Structures

Kai Wörner [Hamburg University]
Andreas Witt [Tübingen University]
Georg Rehm [Tübingen University]
Stefanie Dipper [Potsdam University]
