Combining Markup Semantics and Semantic Markup: A secret marriage

Paper

Sasaki, Felix (University of Bielefeld, Department of Computational Linguistics and Text-technology)
felix.sasaki@uni-bielefeld.de

Contact Information recorded for this paper is:

Felix Sasaki
University of Bielefeld
Faculty for Linguistics and Literature Department of Computational Linguistics
and Text-technology
Postfach 10 01 31
33501 Bielefeld, Germany

Telephone: 0049-(0)521-1062534
FAX: 0049-(0)521-106-2996

The equipment requirements recorded for this paper are:	
Computer Requirements: *(nothing recorded)*	**Special Equipment:** *(nothing recorded)*

Introduction

This proposed paper is concerned with approaches to describe the meaning of markup in (textual) documents, i.e. "markup semantics", and approaches to describe abstract, conceptual models which are sometimes applied to markup, i.e. "semantic markup". A motivation to combine both approaches is discussed, and the respective methodology is explained.

Markup semantics

In the field of humanities computing, textual documents with markup play a crucial role. A sophisticated document grammar like the TEI is a key component for the creation, editing, analysis and interchange of electronic documents. Document grammars describe the syntax of documents. In recent years, a new area of research called "markup semantics" has emerged, which is concerned with the (formal) description of the meaning of these constructs. For example, a <para> element can be used to mark-up paragraphs, but its interpretation as a paragraph is not part of the formal declaration of the element in the document grammar.

The BECHAMEL project (Renear et al., 2002) is probably the most prominent approach in this area of research. Other approaches have been developed (e.g. Simons, 1999; Welty and Ide, 1999) and are discussed in detail elsewhere (Sperberg-McQueen et al., 2000). These approaches share the aim to describe the meaning of markup "bottom-up": Various mechanisms are developed to enhance given

document grammars and marked-up documents with additional, semantic information.

Semantic markup

In contrast to markup semantics, semantic markup[(1)], developed for example within the "semantic web" initiative (Berners-Lee et al., 2001), offers a "top-down" approach, from a given semantic, abstract description of "resources" (a "conceptual level"), to concrete resources (for example to markup). The abstract notion of a "resource" plays a crucial role: the goal is to allow for a meaningful access to every imaginable kind of "resources" on the web. The universality of this approach might become a drawback if it is applied to a specific kind of resource like markup. For example, the "RDF Vocabulary Description Language RDF Schema" (Brickley and Guha, 2003), a central specification in the development of the semantic web, allows to assign properties to classes of resources in the form of triples, e.g `paragraph is-part-of section'. But no mechanism is supplied to verify whether instances of classes really exist and whether they have the properties being assigned, e.g. whether a <para> element is nested inside a <sect> element.

Current research tries to fill this gap, again moving "top-down". Klein et al. , 2001, describe the transformation of abstract, "ontological" descriptions to document grammars. Erdmannn and Studer, 1999, offer a similar approach, focusing on the aspect of querying documents at the conceptual level.

Markup semantics and semantic markup: A unified approach

This proposed paper suggests a combination of the ongoing efforts in the two areas of research described above. Markup is interpreted as a concrete instance of concepts, which are defined at the abstract, conceptual level supplied by the semantic markup approach. A <para> element might be an instance of a concept `paragraph' or of a concept `thematicUnit'. On the other hand, a concept `paragraph' might be instantiated as a <para>, or <absatz> element.

A unified approach to markup semantics and semantic markup offers new prospects for scholars in the humanities: The integration of text encoding and abstract, conceptual resources. Currently, an enormous amount of such resources is being created, for example Metadata relying on the Dublin Core standard, ontological descriptions of linguistic categories (Lewis et al., 2001), or lexical resources like WordNet. These resources can contribute to the solution of some research questions in the field of markup semantics. An example considering multilingual documentation of document grammars and instance documents will be given in the next section.

A "secret marriage": Mapping between markup and the conceptual level

The methodology is visualized in fig. 1. The upper part contains three lexical concepts from the WordNet database, `paragraph', `section' and `book'. They have semantic relations like being a meronym (`paragraph is-meronym-of book'), i.e. a part of book, etc. The lower part of fig. 1. contains markup, i.e. document grammar constructs like the element declaration `<xsd:element name="book"> ...' or instance documents.

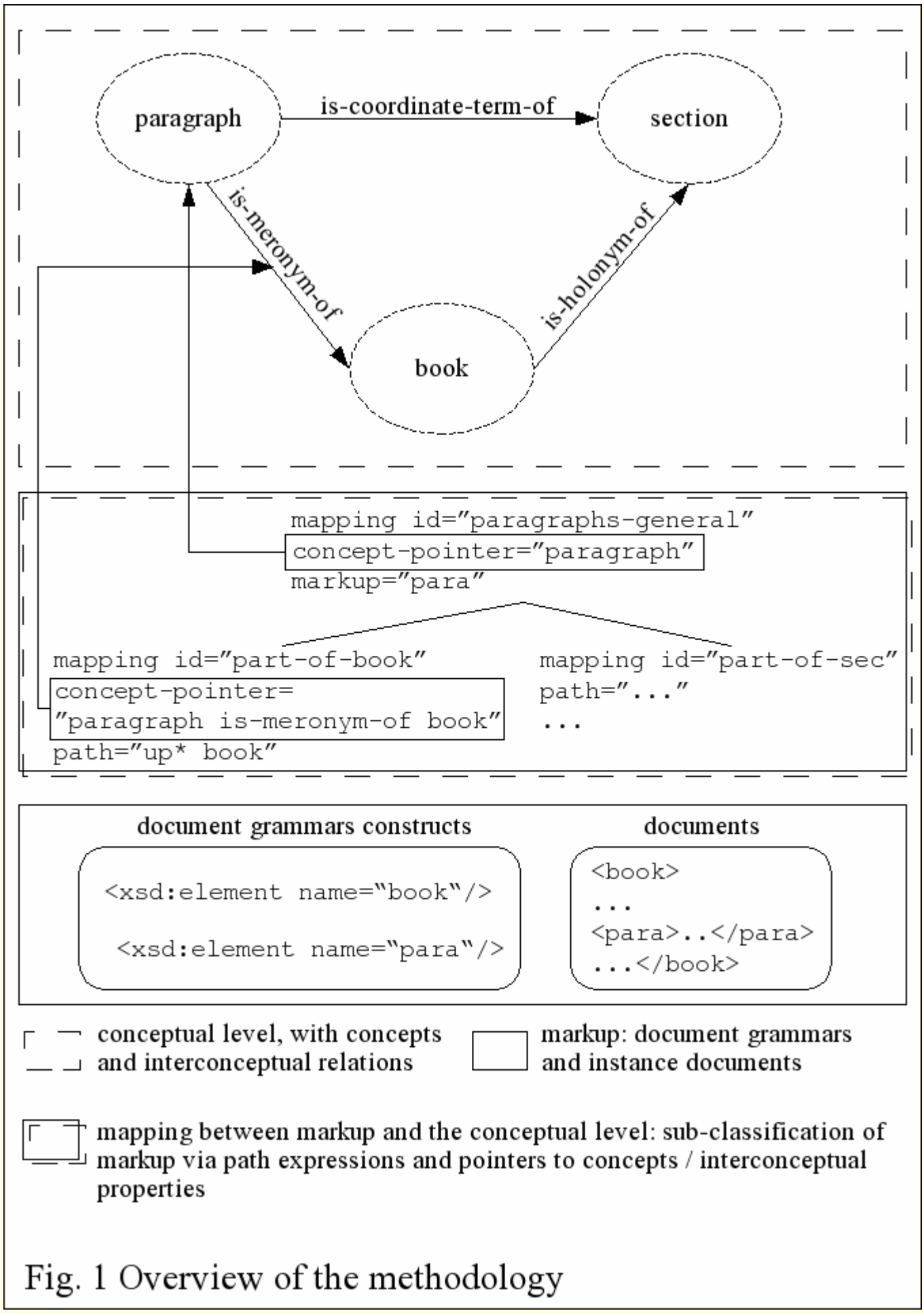

Fig. 1 Overview of the methodology

The key part of the methodology is the mapping between the conceptual level and the markup, which is visualized in the middle of fig. 1. Currently, this mapping is realized within a format called "Context Specification Document" (CSD, Sasaki and P?ninghaus, 2003) [(2)], which allows for the sub-classification

of markup according to its structural properties. The sub-classification is necessary because document grammars do not allow to specify all contextual, structural properties of markup which are necessary for the mapping to the conceptual level. For example in an instance document there might be <para> elements which are nested within a <book> element, or <para> elements which are nested within an <article> element. For the first, a mapping called `part-of-book' is declared. The structural properties of markup are described via path expressions [(3)] like `up* book', which are matched by the respective <para> elements in instance documents. In addition to this specification of structural properties, a CSD contains pointers to concepts or interconceptual, semantic relations. For example the mapping called `part-of-book' has a pointer to the interconceptual relation `paragraph is-meronym-of book'. That is, the meaning of the path expression `up* book' is specified as `paragraph is-meronym-of book'. Using counterparts of WordNet in other languages, e.g. EuroWordNet, a multilingual documentation of markup can be generated, e.g. for German `Absatz ist-Meronym-von Buch'.

Unlike the approaches described above, this mapping can be created "bottom-up" AND "top-down" in a declarative manner: From the conceptual level, a description of meaning can be added to the markup, i.e. as a markup semantics. This allows for the semantic validation of documents during the authoring process. In addition, markup in documents can be retrieved as instances of concepts, i.e. as a semantic markup. For the creation of the mapping and the generation of markup semantics or semantic markup, both markup and the conceptual level do not have to be changed, so this approach is called "a secret marriage".

A prototype of a processor for CSDs, implemented in the programming language Python, creates information about the structural properties found in documents. In addition, XSLT-Stylesheets are being created, to integrate this information into document grammars, into instance documents or into the conceptual level, represented within an XML-serialization of RDF Schema.

Footnotes

1. The terminology "markup semantics" versus "semantic markup" is taken from Renear et al., 2002.

2. An alternative representation for the mapping (Sasaki et al., in press) makes use of RDF to describe other, i.e. non-hierarchical structural properties of markup.

3. For a detailed description of the underlying path language, see Br?gemann-Klein and Wood, 2000.

References

1. Berners-Lee, T., J. Hendler and O. Lassila. The Semantic Web. Scientific American 284 (5), pp. 35-43, 2001.

2. Brickley, D. and R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Working Draft, 2003. http://www.w3.org/TR/rdf-schema/

3. Br?gemann-Klein, A. and D. Wood. Caterpillars: A context specification technique. Markup Languages 1(2), pp. 81-106, 2000.

4. Dublin Core Metadata Initiative. http://dublincore.org/

5. Erdmann, M. and R. Studer. Ontologies as Conceptual Models for XML documents. In: Proceedings of the KAW '99 12th Workshop on Knowledge Acquisition, Modelling and Management. Banff, Canada, 1999.

6. EuroWordNet. http://www.illc.uva.nl/EuroWordNet/

7. Klein, M., D. Fensel, F. van Harmelen and I. Horrocks. The Relation between Ontologies and XML Schemas. In: Electronic Trans. on Artificial Intelligence, Special Issue on the 1st International Workshop "Semantic Web: Models, Architectures and Management". 2001.

8. Lewis W., S. Farrar, D. T. Langendoen. Building a Knowledge Base of Morphosyntactic Terminology. In: Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia, USA, 2001.

9. Renear, A., C. M. Sperberg-McQueen and C. Huitfeldt. Towards a Semantics for XML Markup. In: Proceedings of DocEng' 2002, Virgina, USA, 2002.

10. Sasaki, F. and J. P?ninghaus. Testing Structural Properties in Textual Data: Beyond Document Grammars. Literary and Linguistic Computing 18 (1), pp. 89-100, 2003.

11. Sasaki, F., A. Witt and D. Metzing. Declarations of Relations, Differences and Transformations between Theory-specific Treebanks: A New Methodology. To be published in: Proceedings of the Second Workshop on Treebanks and Linguistic Theories, V?j?University, Schweden, in press.

12. Simons, G. F. Using Architectural Forms to Map TEI Data into an Object-oriented Database. Computers and the Humanities 33, 1(2), pp. 85-101, 1999.

13. Sperberg-McQueen, C. M., A. Renear and C. Huitfeldt. Meaning and Interpretation of Markup. In: Markup Languages: Theory and Practice, 2(3), pp. 215-234, 2000.

14. TEI. Sperberg-McQueen, C. M., and L. Burnard (Eds.). Guidelines for Text Encoding and Interchange (TEI P3). Chicago, Oxford: ACH / ALLC / ACL Text Encoding Initiative, 1994.

15. Welty, C. and N. Ide. Using the Right Tools: Enhancing Retrieval from Marked-up Documents. Computers and the Humanities 33, 1(2), pp. 59-84, 1999.

16. WordNet. http://www.cogsci.princeton.edu/~wn/