Multidimensional markup and heterogeneous linguistic resources

Maik Stührenberg **Andreas Witt** **Daniela Goecke** **Dieter Metzing** **Oliver Schonefeld**
Faculty for Linguistics and Literature
Bielefeld University, Germany
{maik.stuehrenberg|andreas.witt|daniela.goecke|dieter.metzing|oliver.schonefeld}@uni-bielefeld.de

Abstract

The paper discusses two topics: firstly an approach of using multiple layers of annotation is sketched out. Regarding the XML representation this approach is similar to standoff annotation. A second topic is the use of heterogeneous linguistic resources (e.g., XML annotated documents, taggers, lexical nets) as a source for semi-automatic multi-dimensional markup to resolve typical linguistic issues, dealing with anaphora resolution as a case study.[1]

1 Introduction – Why (and how) to use heterogeneous linguistic resources

A large and diverse amount of linguistic resources (audio and video recodings, textual recordings) has been piled up during various projects all over the world. A reasonable subset of these resources consists of machine-readable structured linguistic documents (often XML annotated), dictionaries, grammars or ontologies. Sometimes these are available to the public on the Web, cf. Simons (2004). The availability allows for the sophisticated examination of linguistic questions and the reuse of existing linguistic material. Especially corpora annotated for discourse-related phenomena have become an important source for various linguistic studies. Besides annotated corpora external knowledge bases like lexical nets (e.g., WordNet, GermaNet) and grammars, can be used to support several linguistic processes.

Although XML has recently established itself as the technology and format of choice the before mentioned resources remain heterogeneous in respect of the data format (i.e., the underlying schema) or the functionality provided. A simple approach to make use of different resources is to use them one by one, starting with raw text data (or annotated XML) as input and providing the output of the first process (e.g., a tagger) as input for the next step (e.g., a parser). However, this method may lead to several problems. One possible problem of this method is that the output format of one processing resource can be unemployable as input format for the next. Another potential problem of using XML annotated documents is overlapping annotation. And finally it is sometimes necessary (or desirable) to process only parts of the input document.

The structure of the paper is as follows: In Section 2 our approach of representing multiple annotations is desribed, in Section 3 the use of multi-root trees for the representation of heterogeneous ressources is presented. As a case study, the resolution of anaphoric relations is described in Section 4.

2 Multiple annotations

Representing data corresponding to different levels of annotation is a fundamental problem of text-technological research. Renear et al. (1996) discuss the OHCO-Thesis[2] as one of the basic assumptions about the structure of text and show that this assumption cannot be upheld consistently. Being based on the OHCO-Thesis most markup languages (including SGML and XML) are designed in principle to represent one structure per document. Options to resolve this problem are discussed in Barnard et al. (1995) and several other

[1] The work presented in this paper is part of the project *A2 Sekimo* of the Research Group *Text-technological modelling of information* funded by the German Research Foundation.

[2] The OHCO-Thesis, first presented in DeRose et al. (1990) states "that text is best represented as an ordered hierarchy of content object (OHCO)."

proposals. To avoid drawbacks of the above mentioned approaches Witt (2002; 2004) discusses an XML based solution which is used in our project.

2.1 Representation

We address the issue of overlapping markup by using separate annotations of relevant phenomena (e.g., syntactic information, POS, document structure) according to different document grammars, i.e., the same textual data is encoded several times (in separate files). One advantage of this multiple annotation is that the modeling of information on a level A is not dependent in any way on a level B (in contrast to the standoff annotation model described by Thompson and McKelvie (1997), where a primary modeling level is needed). Additional annotation layers can be added easily without any changes to the layers already established. The primary data (i.e., the text which will be annotated) is separated from the markup and serves as key element for establishing relations between the annotation tiers. Witt et al. (2005) describe a knowledge representation format (in the programming language Prolog[3]) which can be used to draw inferences between separate levels of annotations and in which the parts of text (the `PCDATA` in XML terms) are used as an absolute reference system and as link between the levels of annotation (cf. Bayerl et al. (2003)).

This representation format allows us to use various linguistic resources. A Python script converts the different annotation layers (XML Documents) to the above mentioned Prolog representation which serves as input for the unification process. The elements, attributes and text of all annotation layers are stored as Prolog predicates. As a requirement all files must be identical with respect to their underlying character data, what we call identity of primary data.

2.2 Unification

Figure 1 shows the architecture used for the unification process. Different annotation layers are unified, i.e., merged into one output fact base. A script reconverts the Prolog representation into one well-formed XML document containing the annotations from the layers plus the textual content. In case of overlapping elements (which may be the result of the unification), it converts those elements to milestones or fragments (cf. TEI Guidelines (2004)) according to parameter options. A Java-based GUI is available to ease the use of the above mentioned framework.

3 Multi-rooted trees

Based on the before mentioned architecture we now focus on the usage of heterogeneous linguistic resources (as described in in Section 1) in order to semi-automatically add layers of markup to various XML (or plain text) documents. We prefer the term *multi-rooted trees* in favor of *multiple annotations*, i.e., different layers of annotation are stored in a single representation (based on the above mentioned architecture). The input document is separated into the textual information and the annotation tree (if there is any). Both of these are provided as input for linguistic resources. The output of this process (typically an XML annotated document) is again separated into text and markup and serves as input for another resource. Figure 2 gives an overview over the process.

4 Heterogeneous linguistic resources for anaphora resolution: a case study

We use multi-rooted trees in order to annotate and model coreferential phenomena and anaphoric relations (cf. Sasaki et al (2002)). Base for the resolution of anaphoric relations (both pronominal anaphora and definite description anaphora) is a small test corpus containing German newspaper articles and scientific articles in German and English. Figure 3 shows an excerpt of a German newspaper article taken from "Die Zeit". In this example the first linguistic resource to apply is a parser (in our case Machinese Syntax by Connexor Oy). As a second step an XSLT script uses the input document and the parser output and tags discourse entities (see element `de` in Figure 3) by judging several syntactic criteria provided by the parser. The discourse entities mark the starting point to determine anaphora-antecedent-relations between pairs of discourse entities.

In order to resolve bridging relations (e.g., "door" as a meronym of "room"), WordNet (GermaNet for German texts) is used as a linguistic resource to establish relationships between discourse entities according to the information stored in the synsets[4]. By now, we use an Open Source

[3] An additional representation format integrates the representation of multi-rooted trees developed in the NITE-Project (cf. Carletta et al. (2003)).

[4] A synset represents a concept and consists of a set of one or more synonyms.

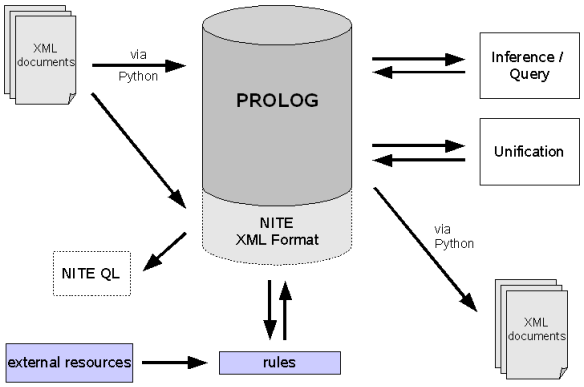

Figure 1: Overview of the architecture

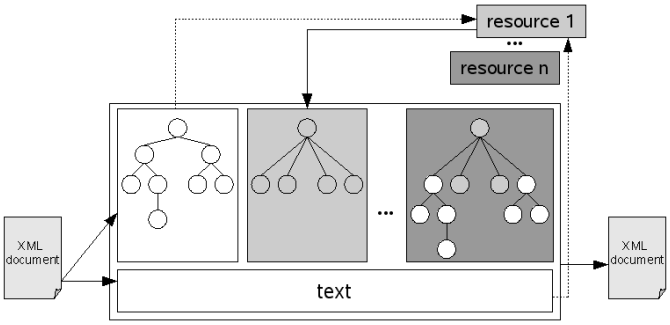

Figure 2: Using heterogeneous linguistic resources

```
<doc_article>
 <doc_sect1>
 <doc_title/>
  <doc_para>
   <chs_sentence id="s22">
    <chs_de deLemma="rolf" deID="de_n_078" deType="nom">Marie Rolfs</chs_de> ist <
     chs_de deLemma="jahr" deID="de_n_079" deType="nom">vier Jahre</chs_de> alt.
   </chs_sentence>
   <chs_sentence id="s23">
    Mit <chs_de deLemma="petra" deID="de_n_080" deType="nom">ihrer
     Zwillingsschwester Petra</chs_de>, <chs_de deLemma="bruder" deID="de_n_081"
     deType="nom">ihrem achtjährigen Bruder</chs_de> und <chs_de deLemma="mutter"
     deID="de_n_082" deType="nom">ihrer Mutter Sabine</chs_de> wohnt <chs_de deLemma
     ="sie" deID="de_n_083" deType="nom">sie</chs_de> in <chs_de deLemma="
     dreizimmerwohnung" deID="de_n_084" deType="nom">einer Dreizimmerwohnung in <
     chs_de deLemma="rahlstedt" deID="de_n_085" deType="nom">Rahlstedt a<chs_de
     deLemma="stadt#rand" deID="de_n_086" deType="nom">m östlichen Hamburger
     Stadtrand</chs_de> </chs_de> </chs_de>.
   </chs_sentence>
   <chs_semRel>
   <chs_cospecLink relType="poss" phorIDRef="de_n_080" antecedentIDRefs="de_n_078"/>
   <chs_cospecLink relType="poss" phorIDRef="de_n_081" antecedentIDRefs="de_n_078"/>
   <chs_cospecLink relType="poss" phorIDRef="de_n_082" antecedentIDRefs="de_n_078"/>
   <chs_cospecLink relType="ident" phorIDRef="de_n_083" antecedentIDRefs="de_n_078"/
    >
   </chs_semRel>
  </doc_para>
 </doc_sect1>
</doc_article>
```

Figure 3: An extract of a German newspaper article marked up by several linguistic resources

native XML database[5] as test tool for querying the GermaNet data[6]. Resolving synonymous or hyperonymous anaphoric relations is done by using XPath or XQuery queries on pairs of discourse entities. Bridging relations are harder to track down and will be focused on in the near future.

Figure 3 shows the shortened and manually revised output of the anaphora resolution. In this example two annotation layers have been merged: the logical document structure (in our case a moddified version of *DocBook*, `doc`) and the level of semantic relations (`chs`). The logical document structure describes the organisation of the text document in terms of chapters, sections, paragraphs, and the like. The level of semantic relations describes discourse entities and relations between them. Corpus investigations give rise to the supposition that the logical text structure influences the search scope of candidates for antecedents. Anaphoric relations are annotated with a `cospecLink` element (lines 12 to 15). The attribute `relType` holds the type of relation between two discourse entities. Line 15 is an example of an identity relation between discourse entity `de_n_078` ("Marie Rolfs", line 6) and discourse entity `de_n_083` ("sie", line 9) whereby the first is marked as antecedent.

5 Conclusion

The architecture shown in this paper provides access to multiple layers of linguistic annotation and allows for the reuse and integration of existing linguistic resources. The resulting additional annotation layers are extremely useful for solving complex linguistic issues like anaphora resolution. It is our goal to enable a semi-automatic anaphora resolution by the end of the project life-span.

References

Barnard, David, Burnard, Lou, Gaspart, Jean-Pierre Gaspart, Price, Lynne A., Spergerg-McQueen, C. M. and Giovanni Battista Varile. 1995. Hierarchical encoding of text: Technical problems and SGML solutions. *Computers and the Humanities*, 29(3):211–231.

Bayerl, Petra Saskia, Lüngen, Harald, Goecke, Daniela, Witt, Andreas and Daniel Naber. 2003. Methods for the semantic analysis of document markup. In *Proceedings of the 2003 ACM symposium on Document engineering*, pages 161–170, New York, NY, USA. ACM Press.

Carletta, Jean Kilgour, Jonathan, O'Donnel, Timothy J., Evert, Stefan 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 (NLPXML-2003))*, Budapest, Hungary.

DeRose, Steven J., Durand, David G., Mylonas, Elli and Allen H. Renear. 1990. What is text, really? *Journal of Computing in Higher Education*, 1(2):3–26.

Renear, Allen, Mylonas, Elli and David Durand. 1996. Refining our notion of what text really is: The problem of overlapping hierarchies. *Research in Humanities Computing. Selected Papers from the ALLC/ACH Conference, Christ Church, Oxford, April 1992*, 4:263–280.

Sasaki, Felix, Wegener, Claudia, Metzing, Dieter Metzing and Jens Pönninghaus. 2002. Co-reference annotation and resources: A multilingual corpus of typologically diverse languages. In *Proceedings of the 3nd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1225–1230, Las Palmas.

Simons, Gary, Lewis, William, Farrar, Scott, Langendoen, Terence, Fitzsimons, Brian and Hector Gonzalez. 2004. The Semantics of Markup: Mapping Legacy Markup Schemas to a Common Semantics. In Graham Wilcock, Nancy Ide and Laurent Romary, editor, *Proceedings of the 4th Workshop on NLP and XML (NLPXML-2004)*, pages 25–32, Barcelona, Spain, July. Association for Computational Linguistics.

Sperberg-McQueen, C. M. and Lou Burnard, editor. 2004. *Guidelines for Text Encoding and Interchange*. published for the TEI Consortium by Humanities Computing Unit, University of Oxford.

Thompson, Henry S. and David McKelvie. 1997. Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe '97: The next decade – Pushing the Envelope*, pages 227–229, Barcelona.

Witt, Andreas, Goecke, Daniela, Sasaki, Felix and Harald Lüngen. 2005. Unification of XML Documents with Concurrent Markup. *Literary and Lingustic Computing*, 20(1):103–116.

Witt, Andreas. 2002. Meaning and interpretation of concurrent markup. In *ALLCACH2002, Joint Conference of the ALLC and ACH*, Tübingen.

Witt, Andreas. 2004. Multiple hierarchies: new aspects of an old solution. In *Proceedings of Extreme Markup Languages*.

[5] eXist (`http://www.exist-db.org`)

[6] GermaNet is available as an XML representation which can be stored (and queried) in the above mentioned (and additional) native XML database.