Multi-Dimensional Markup: N-way relations as a generalisation over possible relations between annotation layers

Harald Lüngen
luengen@uni-giessen.de
Justus-Liebig-Universität, Germany

Andreas Witt
andreas.witt@uni-tuebingen.de
University of Tübingen, Germany

Text-technological background

Multi-dimensional markup is a topic often discussed. The main reasonwhy it is researched is the fact that the most important markup languages today make the implicit assumption that for a document, only a single hierarchy of markup elements needs to be represented. Within the field of Digital Humanities, however, more and more analyses of a text are expressed by means of annotations, and as a consequence, segments of a text are marked up by tags relating to different levels of description. Often, a text is explicitly or implicitly marked up several times. When using the TEI P5 as an annotation scheme one might use markup from different TEI modules concurrently as 'msdescription' for manuscript description, 'textcrit' for Text Criticism, and 'analysis' for (linguistic) analysis and interpretation, because the Guidelines state that "TEI schema may be constructed using any combination of modules" (TEI P5 Guidelines).

Abstracting away from limitations of specific markup languages, textual regions annotated according to different levels of descriptions can stand in various relationsships to each other. Durusau & O'Donnell (2002) list 13 possible relationsships between two elements A and B used to concurrently annotate a text span. Their list comprises the cases 'No overlap' (independence), 'Element A shares end point with start point of element B or the other way round', 'Classic overlap', 'Elements share start point', 'Elements share end point' and 'Element share both their start points and end points'. The latter case is known under the label 'Identity'. The possible relationships between A and B can also be partitioned differently, e.g. into Identity, Region A before region B, or the other way round. Witt (2004) has alternatively grouped the relations into three 'meta-relations' called 'Identity', 'Inclusion', and 'Overlap'. Meta-relations are generalisations over all the 13 basic relations inventorised by Durusau & O'Donnell. The reason for introducing meta-relations is to reduce the number of relations to be analysed to those cases that are most typically needed when querying annotations of multiply annotated documents. The query tool described in Witt et

al. (2005) provides 7 two-way query predicates for the 13 basic relations from Durusau & O'Donnell (where e.g. the two relations overlap(A,B) and overlap(B,A) are handled by one query predicate) and specialised predicates for the three meta-relations.

As argued above, often n-way relationships between elements from three or more annotation layers need to be queried. When the detailed accounts of cases of relations between two elements described above are extended to cases where three or more layers are analysed, the number of possible relationships is subject to a combinatorial explosion and rises into several hundreds and thousands. Only in the case of identity(A,B), additional 13 cases of three-way relationships can be distinguished; for all remaining cases of two-way relationships, considerably more three-way cases need to be distinguished. It seems impossible to invent names, let alone to formulate and implement queries for each one of them. Still, for a user it would be desirable to have a set of query predicates for n-way relations available, lest (s)he needs to repeatedly combine queries for two-way relationships, which often can be done only with the help of a fully-fledged programming language.

Application: Analysing n-way relations in text parsing

One text-technological application where relations between elements on more than two elements need to be analysed, is discourse parsing of argumentative texts. In a bottom-up operating discourse parser such as the one developed for German research articles in the SemDok project (Lüngen et al. 2006), it is checked successively whether a discourse relation holds between two known adjacent discourse segments such that they can be combined to form a larger segment. Often this depends on the presence of a lexical discourse marker, such as the adverb 'lediglich' ('only'), in the second segment. But with 'lediglich' as with numerous other markers, there is the additional condition that it has to occur in the so-called vorfeld (first topological field of a German sentence according to the syntax of German, cf. Hinrichs & Kübler 2006), of the first sentence of the second discourse segment. Thus, a combination of information from at least three different information levels (discourse segments, syntax, and discourse markers) needs to be checked, i.e. whether the following situation holds:

```
L1: <ds>.........................
.........................</ds>
L2: <s><vorfeld>.................
....</vorfeld>.............</s>
L3: <dm>lediglich</dm>
```

This situation corresponds to a meta-relation of three-way inclusion: <ds> from Layer 1 must include a <vorfeld> from Layer 2, which in turn must include a <dm> from Layer 3.

Querying n-way relations between elements of multiple annotations

We have identified a set of n-way meta-relations that are typically needed in text-technological applications for multiply annotated documents, namely N-way independence, N-way identity, N-way inclusion, and N-way overlap, (where independence, identity, and inclusion hold between the elements from all n layers, and overlap holds between at least one pair among the n elements). The proposed poster presentation illustrates further examples from text-technological applications such as discourse analysis and corpus linguistic studies, where querying n-way relations between elements is required. It explains our set of query predicates that have been implemented in Prolog for n-way meta-relations, and how they are applied to the examples. Furthermore it presents an evaluation of their usability and computational performance.

References

Durusau, Patrick and Matthew Brook O'Donnell (2002). *Concurrent Markup for XML Documents*. XML Europe 2002.

Hinrichs, Erhard and Sandra Kübler (2006). What Linguists Always Wanted to Know About German and Did not Know How to Estimate. In Mickael Suominen, Antti Arppe, Anu Airola, Orvokki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari K. Pitkänen and Kaius Sinnemäki (eds.): *A Man of Measure : Festschrift in Honour of Fred Karlsson on his 60th Birthday.* The Linguistic Association of Finland, Special Supplement to SKY Journal of Linguistics 19. Turku, Finland.

Lüngen, Harald, Henning Lobin, Maja Bärenfänger, Mirco Hilbert and Csilla Puskas (2006). Text parsing of a complex genre. In Bob Martens and Milena Dobreva (eds.): *Proceedings of the Conference on Electronic Publishing (ELPUB 2006)*. Bansko, Bulgaria.

Witt, Andreas (2004). Multiple hierarchies: New aspects of an old solution. In *Proceedings of Extreme Markup Languages.* Montreal, Canada.

Witt, Andreas, Harald Lüngen, Daniela Goecke and Felix Sasaki (2005). Unification of XML Documents with Concurrent Markup. *Literary and Linguistic Computing* 20(1), S. 103-116. Oxford, UK.