XML and overlapping hierarchies

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute
Ljubljana
Slovenia

Tsujii Laboratory
University of Tokyo
9.1.2007

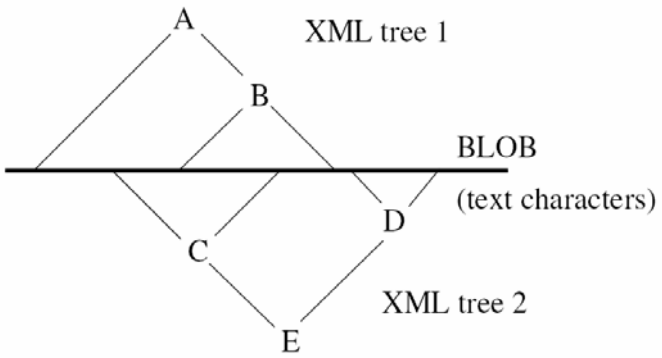
Overview

1. **the problem**
2. **in-line methods**
   1. **fragmented markup**
   2. **milestones**
3. **stand-off methods**
   1. **multiple annotation**
   2. **using pointers**
4. **GENIA**

The OHCO concept

- Text as Ordered Hierarchy of Content Objects (OHCO)
- ..each structure properly nests within the higher level one
- XML is a tree-modelling language - well suited for OHCO representations
- however, not all structures are tree-like..

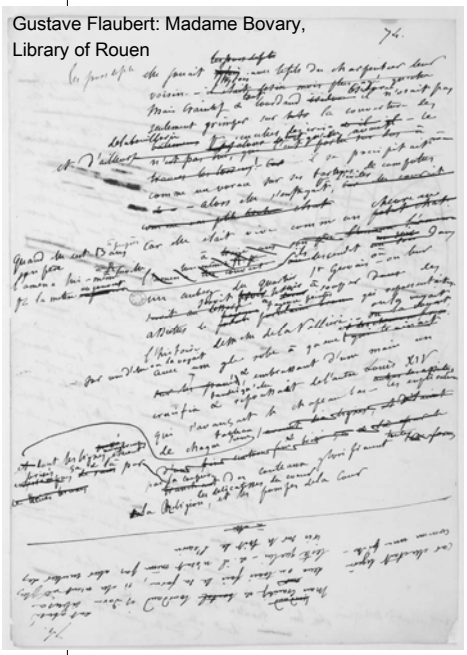
Main problem: crossing hierarchies



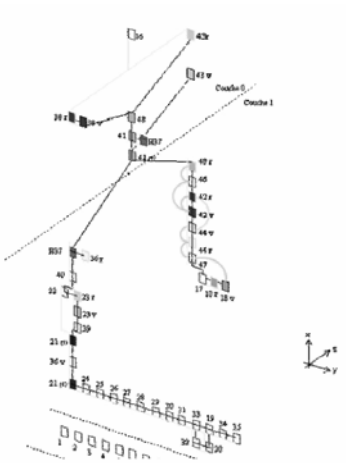
Problems in Humanities research ~ text modelling

- layout vs. paragraph structure:
  <page> … … </page> …
- paragraph vs. reported speech:
   … <q> … … </q>
- metrical vs. syntactic structure:
  <l><s>….</l>….</s>
- etc.
- research in the humanities - mostly done in the domain of so called text-critical editions and transcriptions of primary sources

Gustave Flaubert: Madame Bovary, Library of Rouen



A complex case



Jean-Daniel Fekete
IN|SITU| Project
INRIA Futurs, France
http://insitu.lri.fr/~fekete/

Many proposals!

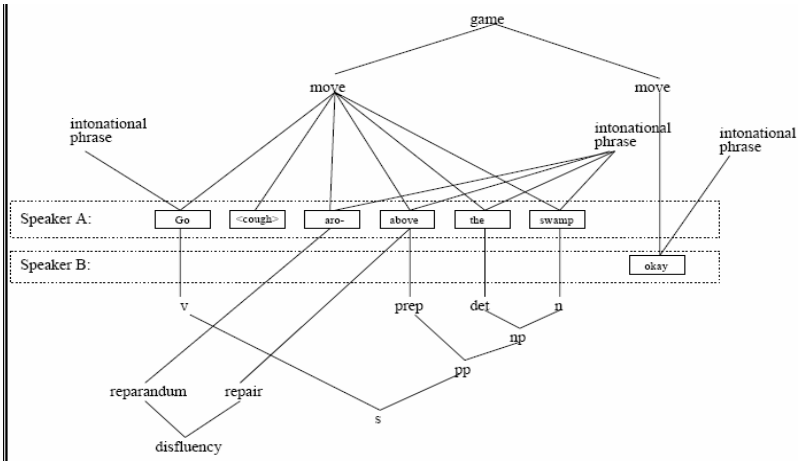
From: C. M. Sperberg-McQueen. Rabbit/duck grammars: a validation method for overlapping structures. *Extreme Markup Languages 2006*. Montréal, Québec:

…ways of dealing with non-hierarchical information; see, for example, [Barnard et al. 1988], [Renear et al. 1993], [Barnard et al. 1995], [Huitfeldt 1995], [Murata 1995], [Durand et al. 1996]. Out-of-line or standoff markup [Dybkjær et al. 1998], fragmentation of elements, milestone elements, virtual elements (e.g. [Barnard et al. 1988], [ACH/ACL/ALLC 1994], [Barnard et al. 1995]), concurrent structures [ISO 1986] [Sperberg-McQueen / Huitfeldt 1999], MECS [Huitfeldt 1999], parallel encoding (e.g. [Witt 2004], [Hilbert et al. 2005], [Dekhtyar / Iacob 2005]), bottom-up virtual hierarchies [Durusau / O'Donnell 2002a], layered markup and annotation (LMNL) [Tennison / Piez 2002], range algebra [Nicol 2002], just-in-time trees [Durusau / O'Donnell 2002b], multi-colored trees [Jagadish et al. 2004], tables [Durusau / O'Donnell 2004], TexMecs [Huitfeldt / Sperberg-McQueen 2001], Goddag structures [Sperberg-McQueen / Huitfeldt 2000], Trojan horses [DeRose 2004]

Problems in NLP ~ corpus annotation

- probably the most work in this area has been done for annotation of speech corpora, dialogues and discourse structure
- crossing hierarchies:
  phonological / morphemic; discourse / syntactic
- discontinuous constituents:
  syntax, multi-word expressions, disfluencies
  - note that in NLP a similar problem occurs with parsing
- also: annotation of same level by different tools

Example: dialogue vs. syntactic annotation



from: Jean Carletta, David McKelvie, and Amy Isard: **Supporting linguistic annotation using XML and stylesheets. In** *Readings in Corpus Linguistics*, ed. G. Sampson and D. McCarthy, London and NY: Continuum International, 2002.

Solutions

- In-line markup:
  - modified markup is included in the text file
  - main advantage: text / markup can still be (partially) hand edited and validated
  - favoured by the humanities community
  - main standard used:
    - TEI
- Stand-off markup:
  - markup is stored separately, and points to text
  - main advantage: any relationship can be expressed
  - favoured by the NLP community
  - main standard (beginning to be) used:
    - ISO TC37 SC4 proposals

Types of overlap

(adapted from: Patrick Durusau, Matthew Brook O'Donnell. Concurrent Markup for XML Documents. XML Europe 2002)

In-line solutions

- standard approaches:
  - fragmented markup
  - milestones
- alternatives: BUVH, etc.
- non-XML approaches:
  - TexMECS
  - JITT
  - LMNL

Fragmented markup

- One hierarchy taken as primary, for the other(s) the markup is modified so that each “crossing” element is split
- Non-well formed structure:
  <s>According to the visiting leader, the economy of the country is <q>"better than ever.</s><s>It is in fact in very good shape.</s>"</q>
- With fragmented <q>:
  <s>According to the visiting leader, the economy of the country is <q id=“q1” next=“q2”>"better than ever.</q></s><q id=“q2” prev=“q1”><s>It is in fact in very good shape.</s>"</q>

Fragmented markup II

- Advantages:
  - retains as much as possible of the structure
  - XML schema need not be changed
  - not very difficult to implement
- Disadvantages:
  - one hierarchy arbitrarily chosen as the primary one
  - needs special (XSLT, XQuery) constructs to implement
  - elements have be broken up into many pieces if:
    - secondary hierarchy crosses multiple branches of primary one
    - we have to deal with multiple hierarchies

Milestones

- Container elements are substituted by empty elements
- Example:
  - Non-well formed structure:
    <s>According to the visiting leader, the economy of the country is <q>"better than ever.</s><s>It is in fact in very good shape.</s>"</q>
  - With milestone <q>:
    <s>According to the visiting leader, the economy of the country is <q-start id="q1" coid="q2"/>"better than ever. </s><s>It is in fact in very good shape.</s>"</q-end id="q2" coid="q1">
- Choices: only broken or all elements can be milestoned
- Best type of milestones, so called Troyan milestones used by OSIS:
  - use <q who='paris'>...</q> when you can, otherwise
  - use <q who='paris' sID='foo'/>...<q eID='foo'/>

Milestones II

- Advantages:
  - conceptually simple
  - can express any type of overlap
  - simple to change to/from other formats
- Disadvantages:
  - retain no structure in the document - very hard or impossible to do validation or querying
  - have to (substantially) change the schema

Non-XML approaches

- These approaches introduce new syntax and semantics
- best known probably MECS / TexMECS:
  Huitfeldt, Claus, and C. M. Sperberg-McQueen. 2001. "TexMECS: An experimental markup meta-language for complex documents".
  http://helmer.aksis.uib.no/claus/mlcd/papers/texmecs.html
- example:
  {sp{{speaker{AASE}speaker}{l{Peer, you're lying!}-l}}sp}
  {sp{{speaker{PEER GYNT }speaker}
  {stage{without stopping}stage}{+l{No, I'm not!}l}}sp}
  {sp{{speaker{AASE}speaker}{l{Well then, swear to me it's true.}l}}sp}
  {sp{{speaker{PEER GYNT}speaker}{l{Swear? why should I?}-l}}sp}
  {sp{{speaker{AASE}speaker}{+l{See, you dare not!}l} {l{Every word of it's a lie.}l}}sp}

XML in-line solutions - conclusions

- it is still possible to modify text and (partially) markup
- partial XML validation still possible, although schema might need to be changed
- smaller or greater problems with using standard methods of extracting information (XPath, XQuery)
- become less advantageous as number of different levels grow
- we might not want to deal with all the markup all the time

Stand-off solutions

- Markup is separated from text
- markup points to text using XLink
- (dis)advantages: any type of relationships possible
- Methods discussed:
  - hybrid approaches:
    - joins
    - multiple annotations
  - pure stand-off:
    - pointers (& tools)
    - RDF

Joins

```
<sp who="hughie"><p>How does it go?
  <q><l id="x1">da-da-da</l>
     <l id="l2">gets a new frog</l>
     <l>...</l>
  </q></p></sp>
<sp who="louie"><p><q><l id="l1">When the old pond</l> ...</q></p></sp>
<sp who="dewey"><p><q>... <l id="l3">It's a new pond.</l></q></p>
<!-- ... -->
<join targets="l1 l2 l3" result="lg" desc="haiku" scope="root"/>
</sp></para>
```

- “semi stand-off markup”
- introduced by TEI
- advantage: power - any relationship can be expressed
- disadvantage: join object is not (in general) contiguous with the segments it is joining

Join-like syntactic annotation

Used in the TIGER treebank of German

```
<s id="s4231">
 <graph root="s4231_VROOT" discontinuous="true">
  <terminals>
   <t id="s4231_1" word="In" lemma="in" pos="APPR" morph="--"/>
   <t id="s4231_2" word="Japan" lemma="Japan" pos="NE" morph="Dat.Sg.Neut"/>
   <t id="s4231_3" word="wird" lemma="werden" pos="VAFIN" morph="3.Sg.Pres.Ind"/>
   <t id="s4231_4" word="offenbar" lemma="offenbar" pos="ADJD" morph="Pos"/>
   <t id="s4231_5" word="die" lemma="der" pos="ART" morph="Nom.Sg.Fem"/>
   <t id="s4231_6" word="Fusion" lemma="Fusion" pos="NN" morph="Nom.Sg.Fem"/>
   <!-- ... -->
   <t id="s4231_18" word="." lemma="--" pos="$." morph="--"/>
  </terminals>
  <nonterminals>
   <nt id="s4231_500" cat="PP">
    <edge label="AC" idref="s4231_1"/>
    <edge label="NK" idref="s4231_2"/>
   </nt>
   <nt id="s4231_501" cat="CNP">
    <edge label="CJ" idref="s4231_9"/>
    <edge label="CD" idref="s4231_10"/>
    <edge label="CJ" idref="s4231_11"/>
   </nt>
   <!-- ... -->
   <nt id="s4231_507" cat="S">
    <edge label="OC" idref="s4231_504"/>
    <edge label="HD" idref="s4231_3"/>
    <edge label="SB" idref="s4231_506"/>
   </nt>
   <nt id="s4231_VROOT" cat="VROOT">
    <edge label="--" idref="s4231_507"/>
    <edge label="--" idref="s4231_18"/>
   </nt>
  </nonterminals>
 </graph>
```


Multiple Annotation

- Developed by Andreas Witt, Bielefeld University, c.f. http://www.text-technology.de/, Sekimo project
- each annotation layer is a separate XML document, which contains both markup and text
- the text serves as the implicit link
- text has two representations:
  - as XML documents
  - as a Prolog database
- they can be programmatically derived and used together for editing, inference, or unification of the multiply annotated document
- advantages:
  - each level can be viewed separately
  - simple to add new levels

Multiple Annotations II

Prolog representation for elements:

```
node('d-xhtml.xml',729,786,[1,5,3,2],element('td')).
node('d-xhtml.xml',729,786,[1,5,3,2,1],element('ul')).
node('d-xhtml.xml',729,751,[1,5,3,2,1,1],element('li')).
node('d-thema.xml',729,786,[1,5,3,2],element('causes')).
node('d-thema.xml',729,751,[1,5,3,2,1],element('cause')).
```

Similar structure for attributes:

```
attr('tape-xhtml.xml',729,786,[1,5,3,2],'valign','top').
```

For PCDATA:

```
pcdata_node(729, 730, 'N').
pcdata_node(730, 731, 'o').
pcdata_node(731, 732, ' ').
pcdata_node(732, 733, 'c').
pcdata_node(733, 734, 'a').
pcdata_node(734, 735, 't').
```

(adapted from *Andreas Witt.* Multiple hierarchies: new aspects of an old solution. Extreme Markup Languages 2004, Montréal, Québec)

Multiple Annotations III

Software support:

- `xml2prolog.py`
  XML documents to Prolog
- `nexus.pl`
  Prolog to NITE format
- `semt.pl`
  Prolog to XML (milestone/fragmentation of incompatible elements)
- markup can be modified with standard programs
- text can be edited with 2 specialised editors
- programs are freely available
- performance might be an issue

“Proper” pointers

```
<link xlink:type='extended'>
   <anchor xlink:type='locator' xlink:role='quote' href="#l1">
   <anchor xlink:type='locator' xlink:role='quote' href="#l2">
   <anchor xlink:type='locator' xlink:role='quote' href="#l3">
</link>
```

- used by most NLP approaches, also for “proper” hierarchies
- no (or only basic) markup present in text
- markup separated into one or several distinct files
- XPointer/XLink attributes to point to IDs or simply text offsets

Pointers II

- advantages: power - any relationship can be expressed
- disadvantages:
  - very difficult to perform validation
  - (difficult to query the data)
  - difficult to change the annotations
  - impossible to change the text
- this approach best for “read only” systems: annotations and esp. text are fixed, the only interest is in accessing the data
- still, lots of annotation tools exist that know how to deal with stand-off annotations

Tool: MonetDB/XQuery

Alink et al. Representing and Querying Multi-dimensional Markup for Question Answering. NLPXML-2006, Trento, Italy

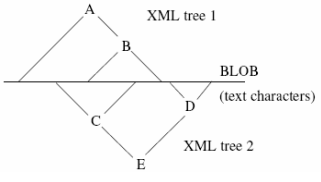
- supports overlapping markup
- source text, or character data, is stored as a Binary Large OBject (BLOB)
- all annotations stored in a single XML document, pointing to BLOB with offsets
- querying performed with extended XQuery
- system, reported to be very fast and scalable, is implemented in a special (open source) DB MonetDB/XQuery

MonetDB/XQuery II

Main innovation: 4 new XPath axes

- e.g. XML document:

```
<A start="10" end="50">
  <B start="30" end="50"/>
</A>
<E start="20" end="60">
  <C start="20" end="40"/>
  <D start="55" end="60">
</E>
```



- //B/select-wide::*
  returns all nodes that overlap with the span of a B node:
  in our case A, B, C and E.
- //*[./select-narrow::B]
- returns nodes that contain the span of B:
  in our case, A and E.

Context	Axis	Result nodes
A	select-narrow	B C
A	select-wide	B C E
A	reject-narrow	E D
A	reject-wide	D

Table 1: Example annotations.

Tool: NITE

http://www.ltg.ed.ac.uk/NITE

- NITE XML Toolkit (NXT) is open source Java software for working with multimodal, spoken, or text language corpora
- designed to support the tasks of human annotators and analysts of heavily cross-annotated data sets
- allows combination of multiple audio and video signals with crossing structures of linguistic annotation
- has been already used on a range of projects with varying needs

NITE and crossing hierarchies

- NITE uses its internal data representation based on multi-rooted trees: nodes can have one set of children, but multiple parents from different upward trees
- data is serialized into XML by dividing the multi-rooted tree into convenient trees where the XML structure mirrors the data structure and representing the remaining connections between nodes using stand-off links with XLink
- NXT Search: stand-alone program to query NITE-type annotated corpora
  - uses special syntax for querying

Other tools

Quite a few other (well-known) tools use stand-off markup:

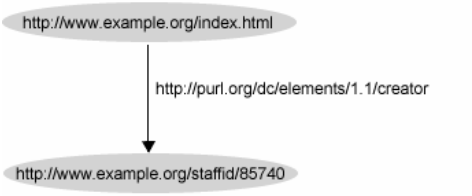
- Callisto,
- MMAX,
- AGTK (Annotation Graph Toolkit)
- ATLAS (Architecture and Tools for Linguistic Analysis Systems)
- Wordfreak
- etc.

Radical Stand-off: RDF

- The Resource Description Framework is a language intended for representing metadata about Web resources.
- RDF identifies things using URIs
- RDF describes resources in terms of simple properties and property values
- RDF statements can be represented as a graph of nodes and arcs representing the resources, their properties and values
- RDF provides an XML-based syntax (called RDF/XML) for recording and exchanging these graphs
- RDF is the main language of the Semantic Web, e.g. it is the basis of OWL, the Ontology Web Language

Simple RDF example

- Example of a statement about a Web page: *http://www.example.org/index.html* has a *creator* whose value is *John Smith*
- RDF terms for the 3 parts of these statement are:
  - **subject**: *http://www.example.org/index.html*
  - **predicate**: *creator*
  - **object**: *John Smith*
- All 3 are **URI references**, e.g.
  - subject URI, e.g. *http://www.example.org/index.html*
  - predicate URI, e.g. *http://purl.org/dc/elements/1.1/creator*
  - object URI, e.g. *http://www.example.org/staffid/85740*
- RDF graph:



Using RDF for linguistic annotation

Aguado de Cea et al. RDF(S)/XML linguistic annotation of semantic web pages. *2nd workshop on NLP and XML*, 2002.

- example gives morphosyntactic annotations with three taggers
- they also describe syntactic and semantic annotation
- however, no tools for querying such structures are presented

```
<contentWeb:FilmReview>
    <contentWeb:text>Tras cinco años de espera y después de
          muchas habladurías, llega a nuestras pantallas la película
          más esperada de los últimos tiempos.</contentWeb:text>
</contentWeb:FilmReview>

<!-- Morpho-syntactic annotation excerpt -->

<morphAnnot:Word rdf:ID="1_16">
    <morphAnnot:surface_form>la</morphAnnot:surface_form>
    <morphAnnot:TradAnnot rdf:about="#trad_ann_info_1_16"/>
    <morphAnnot:MBTAnnot rdf:about="#mbt_ann_info_1_16"/>
    <morphAnnot:ConstrAnnot rdf:about="#constr_ann_info_1_16"/>
</morphAnnot:Word>

<morphAnnot:TradAnnot rdf:ID="trad_ann_info_1_16">
    <trad:tag> ARTDFS </trad:tag>
    <morphAnnot:lemma> el </morphAnnot:lemma>
</morphAnnot:TradAnnot>

<morphAnnot:MBTAnnot rdf:ID="mbt_ann_info_1_16">
    <mbt:tag> TDFS0 </mbt:tag>
    <morphAnnot:lemma> el </morphAnnot:lemma>
</morphAnnot:MBTAnnot>

<morphAnnot:ConstrAnnot rdf:ID="constr_ann_info_1_16">
    <constr:tag> DET </constr:tag>
          <constr:genus>FEM</constr:genus>
          <constr:numerus>SG</constr:numerus>
    <morphAnnot:lemma>la</morphAnnot:lemma>
    <constr:synfunction>DN&gt;</constr:synfunction>
</morphAnnot:ConstrAnnot>
```

Annotea

- W3C project for annotating Web pages
- uses Annotea server, where people can deposit notes / comments on regions of Web pages
- uses RDF
- Web pages + annotations can be viewed with an Annotea aware Web browser
- Basic idea similar to corpus annotation
- But annotations much less dense, and more dispersed

Why is RDF not more popular?

- Still a big divide between the Semantic Web and the NLP communities?
- Data structures are very large, larger than is necessary even with stand-off annotation
- No tools yet exist that would support RDF corpus annotations

Stand-off solutions - conclusions

- can represent any relationship between structures
- difficult to use standard XML methods (XPath, XQuery) to extract information
- difficult to validate structures
- difficult to manually modify markup
- impossible to modify text

→ useful only with specialised tools to operationalise such markup

Overlapping Markup Desiderata

Steven DeRose. Markup Overlap: A Review and a Horse in Extreme Markup Languages 2004 (Montréal, Québec)

- Adequacy
- Human readability
- Maintainability
- Available implementations
- XML compatibility
- Ease of validation
- Validation across hierarchies
- Ease of formatting
- Ease of extracting multiple views
- Ease of extracting hierarchical subsets
- Continuity of text content

GENIA

- several copies of the corpus, each annotated with different linguistic information
- at certain points the text itself differs
- how to merge the separate copies into one corpus?
- how to query over all / subsets of the annotations?
- how to ensure easy addition of further annotations?

Conclusions

- Many different approaches, with different strengths and weaknesses
- In-line approaches can - to an extent - rely on existing XML technologies
- Stand-off approaches need special tools
  - however, quite a few exist
- For GENIA, maybe:
  - Hand editing of text:
    Milestones
  - Maintaining separate copies:
    Multiple annotations
  - Fast and flexible querying:
    Monet/XQuery