Automated Markup Tools

TEI@Oxford

July 2009
Automating the markup process

- People don't like typing pointy brackets...
- ... so they do it wrong, or inconsistently!
- What tools can help us automate this process?
  - general transformation tools, e.g. perl
  - word-processor add-ons e.g. for Open Office
  - specifically XML tools, e.g. XSLTPROC, EGE
  - tools for linguistic annotation
  - tools for image annotation
“Plain Vanilla ASCII”

- Search and replace techniques will usually capture
  - Paragraph structure (blank lines)
  - Headings (lines in caps)
  - (sometimes) emphasis (strings between _ or *)

- Watch out for
  - Markup characters in the text
  - Metadata information

- Use:
  - Your favourite editor
  - Perl
  - (once you are well-formed) xslt transforms
Simple example

In a plain text file, which is *completely* regular, we can even write an XSLT stylesheet to do the job!

1. Make the plain text into well-formed XML
2. Use XSLT to enhance its markup
VARNEY, THE VAMPYRE;

or,

THE FEAST OF BLOOD.

P R E F A C E .

-------

The unprecedented success of the romance of "Varney the Vampyre," leave the Author but little to say further, than that he accepts that success and its results as gratefully as it is possible for any one to do popular favours.

A belief in the existence of Vampyres first took its rise in Norway and Sweden, from whence it rapidly spread to more southern regions, taking a firm hold of the imaginations of the more credulous portion of mankind.

The following romance is collected from seemingly the most authentic sources, and the Author must leave the question of credibility entirely to his readers, not even thinking that he his peculiarly called upon to express his own opinion upon the subject.

Nothing has been omitted in the life of the unhappy Varney, which could tend to throw a light upon his most extraordinary career, and the fact of his death just as it is here related, made a great noise at the time through Europe, and is to be found in the public prints for the year 1713.

With these few observations, the Author and Publisher, are well content to leave the work in the hands of the public, which has stamped it with an approbation far exceeding their most sanguine expectations, and which is calculated to act as the strongest possible ncentive to the production of other works, which in a like, or perchance a still further degree may be deserving of public patronage and support.

To the whole of the Metropolitan Press for their laudatory notices, the Author is peculiarly obliged.

_London Sep._ 1847

+-+

+-----------------------------------------+
|   This Varney the Vampyre e-text was entered by members of the     |
+-----------------------------------------+
Varney: the stylesheet

```xml
<xsl:template name="doParas">
  <xsl:param name="s" select="."/>
  
  <xsl;element name="p">
    <xsl:choose>
      <xsl:when test="contains($s,$blankLine)">
        <xsl:value-of select="substring-before($s,$blankLine)"/>
      </xsl:when>
      <xsl:otherwise>
        <xsl:value-of select="$s"/>
      </xsl:otherwise>
    </xsl:choose>
  </xsl:element>
</xsl:template>
```

VARNEY, THE VAMRYE;
or,
THE FEAST OF BLOOD.

P R E F A C E.

---------

The unprecedented success of the romance of "Varney the Vampyre," leave the Author but little to say further, than that he accepts that success and its results as gratefully as it is possible for any one to do popular favours.

A belief in the existence of Vampyres first took its rise in Norway and Sweden, from whence it rapidly spread to more southern regions, taking a firm hold of the imaginations of the more credulous portion of mankind.

The following romance is collected from seemingly the most authentic sources, and the Author must leave the question of credibility entirely to his readers, not even thinking that he his peculiarly called upon to express his own opinion upon the subject.

Nothing has been omitted in the life of the unhappy Varney, which could tend to throw a light upon his most extraordinary career, and the fact of his death just as it is here related, made a great noise at the time through Europe, and is to be found in the public prints for the year 1713.

With these few observations, the Author and Publisher, are well content to leave the work in the hands of the public, which has stamped it with an approbation far exceeding their most sanguine expectations, and which is calculated to act as the strongest possible ncentive to the production of other works, which in a like, or perchance a still further degree may be deserving of public patronage and support.

To the whole of the Metropolitan Press for their laudatory notices, the Author is peculiarly obliged.

London Sep., 1847
Silk purses and sows ears

- TEI is an XML vocabulary
- Word and Open Office both use XML vocabularies
- so converting one to the other is a simple XSLT transformation, right?

Up to a point, Lord Copper
Converting from Word DOC to TEI XML

One simple way is to use the TEI XML filter supplied for Open Office, (teioop5.jar), downloadable from Sourceforge

- Install the jar file in your copy of Open Office
- Open the Word file
- Choose Save As and select TEI P5 from the list
- Use XSLT to improve on the tagging you get back
Word to TEI example (1)

To download files, click on the "Word" or "text" links below. All files are compressed in ZIP format. The Old Paths Baptist Institute of the Bible web site. Please contact me if you have any difficulties.

<table>
<thead>
<tr>
<th>Downloads</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>entire KJV Bible, single large file</td>
<td>Word, 1.5MB</td>
</tr>
<tr>
<td>entire KJV Bible, folder with individual files</td>
<td>Word, 1.8MB</td>
</tr>
<tr>
<td><strong>1611 Preface to the Reader</strong></td>
<td>Word, 36KB</td>
</tr>
<tr>
<td><strong>1611 Epistle Dedicatory</strong></td>
<td>Word, 8KB</td>
</tr>
<tr>
<td>Genesis</td>
<td>Word, 88KB</td>
</tr>
<tr>
<td>Exodus</td>
<td>Word, 80KB</td>
</tr>
<tr>
<td>Leviticus</td>
<td>Word, 48KB</td>
</tr>
<tr>
<td>Numbers</td>
<td>Word, 64KB</td>
</tr>
<tr>
<td>Deuteronomy</td>
<td>Word, 60KB</td>
</tr>
<tr>
<td>Joshua</td>
<td>Word, 44KB</td>
</tr>
<tr>
<td>Judges</td>
<td>Word, 40KB</td>
</tr>
<tr>
<td>Ruth</td>
<td>Word, 12KB</td>
</tr>
</tbody>
</table>
The Book of Ruth

Chapter 1

1 Now it came to pass in the days when the judges ruled, that there was a famine in the land. And a certain man of Bethlehemsaid to his wife, Elimelech, I will go to the country of Moab to sojourn there, for the famine was grievous in the land.

2 And the husband of the woman was Elimelech, and the name of his wife Naomi, and the name of his two sons Mahlon and Chilion, of the children of Bethlehem of Judah. And they came into the country of Moab, and continued there.

3 And Elimelech, Naomi's husband, died, and she was left, and her two sons.

4 And they took them wives of the women of Moab: the name of the one was Orpah, and the name of the other Ruth. And they dwelt there about ten years.

5 And Mahlon and Chilion also died both of them; and she was left of her two sons and her husband.

6 Then she came with her daughters-in-law from the country of Moab, for she had heard that the LORD had visited his people in Moab; for she had heard how that the LORD had visited his people in Moab.}

for I am too old to have an husband. If I should say, I have hope, if I should have an husband also to night, and should also be as a woman.

13 Would ye tarry for them all they were grown? would ye stay for them from having husbands? say, my daughter, go, I pray thee, to thy mother's house, and abide there: for why should I cast my bread? for why should I set my table for them, while the young men of Moab say, is not the Hebrew woman better than the Moabitish woman? 14 And they lifted up their voice, and wept again, and Orpah and Ruth kissed their mother-in-law, and they returned to their own country.

15 And she said, Behold, thy sister-in-law is gone back unto her people, and unto her gods: return thou after thy sister-in-law.

16 And Ruth said, Intermitt I will not leave thee, or to return from following after thee: whether thou goest, I will go; and where thou lodgest I will lodge: thy people shall be my people, and thy God my God.

17 Where thou shalt die will I die, and there will I be buried: the LORD do so to me, and more also, if ought but death part thee from me.

18 When she saw that she was determined to go with her, then she left speaking with her.
<revisionDesc>
  <change>
    <name>Bruce Wilcox</name>
    <date>2003-11-13T22:53:00</date>
  </change>
</revisionDesc>
</teiHeader>
<text>
  <body>
    <p>THE BOOK OF</p>
    <p>RUTH</p>
    <p>CHAPTER 1</p>
    <p><hi>1</hi> Now it came to pass in the days when the judges re</p>
    <p><hi>2</hi> And the name of the man <emph>was</emph> Elimelech, and the name of his</p>
    <p><hi>3</hi> And Elimelech Naomi's husband died; and she was left, and her two sons.</p>
    <p><hi>4</hi> And they took them wives of the women of Moab; the</p>
    <p><hi>5</hi> And Mahlon and Chilion died also both of them; and the woman was left of her two sons and her husband.</p>
    <p><hi>6</hi> ¶ Then she arose with her daughters in law, that she might return from the country of Moab: for she had heard in the coun</p>
  </body>
</text>
A case study: adding POS tagging to Punch

• Q. What do we mean by POS tagging?
• A. Something like this:
  http://ucrel.lancs.ac.uk/claws/trial.html
• See http://www-nlp.stanford.edu/links/statnlp.html#Taggers for a long list of such things.

We like treetagger...
IPP annotation workflow

For each file,

1. extract the `<text>` element only and pass it through treetagger
2. post-process the treetagger output to make it well-formed XML
3. combine that with the header of the file
<div type="div2">
  <head>RESOLUTIONS.</head>
  <lg>
    <l>I will not breakfast in my bed</l>
    <l>With downy cushions at my head</l>
    <l>That would be very wrong—and so</l>
    <l>Away the eggs and bacon go</l>
  </lg>
  <lg>
    <l>I will not read in bed at night</l>
    <l>And burn the dear electric light</l>
    <l>Nor buy another costly hat</l>
    <l>Oh no! I'm much too good for that</l>
  </lg>
  <lg>
    <l>But I will rise before the dawn</l>
    <l>And weed and cut and roll the lawn</l>
    <l>My border I will plant with veg</l>
    <l>Abundantly from hedge to hedge</l>
  </lg>
  <lg>
    <l>And all the day I'll practise thrift</l>
    <l>And no more happily will drift</l>
    <l>In deeper debt, as once, alas</l>
    <l>But what an awful year I'll pass</l>
  </lg>
  <milestone rend="hr" unit="rule"/>
</div>
I will not break breakfast in my bed with downy cushions at my head; that would be very wrong—and away the eggs and bacon!
IPP-POS example -3

<div type="div2">
  <head>
    <s n="1">
      <w type="NNS" lemma="resolution">RESOLUTIONS</w>
      <c type="SENT">.</c>
    </s>
  </head>
  <lg>
    <s n="2">
      <w type="I">I</w>
      <w type="will">will</w>
      <w type="not">not</w>
      <w type="breakfast">breakfast</w>
      <w type="in">in</w>
      <w type="my">my</w>
      <w type="bed">bed</w>
      <lb/>
      <w type="With">With</w>
      <w type="downy">downy</w>
      <w type="cushion">cushions</w>
      <w type="at">at</w>
      <w type="my">my</w>
      <w type="head">head</w>
      <lb/>
      <w type="That">That</w>
      <w type="would">would</w>
      <w type="be">be</w>
      <w type="very">very</w>
      <w type="XXXX">wrong–and</w>
      <w type="so">so</w>
      <lb/>
    </s>
  </lg>
</div>