CHALLENGES IN CAPTURING AND MODELLING HUMANITIES DATA

Owen Conlan and Seamus Lawless

CNGL, School of Computer Science and Statistics, Trinity College Dublin
3 CHALLENGES

• Consistency
• Authority
• Representation
• Involves Human and Machine processes

• Human expert: transcribe, interpret, markup etc.

• Machine techniques: harmonise, normalise, uplift entities, linguistically analyse, etc.

• Trade-off between completeness and time
1 CORPUS / 3 PROJECTS

• 1641 Depositions Project

• Language and Linguistic Evidence in the 1641 Depositions

• CULTURA
• 31 volumes
• ~8,000 depositions
• 19,010 pages
• ~3.5 million words
“that one whoe was a Scochman pinched with extreame hunger privately in the night tyme opened the grave of a man that was buried within the liberties of the Castle and fed vpon the dead and buried mans flesh Some of the souldjers of the Castle partly espyring him tooke ayme & thincking him an enemy shott him through soe that he dyed, And that [Marginalia:symbol] Scochmans wife afterwards hanged to death her owne child and eate her flesh for want of meate”

MS 814, County Offaly (King’s County)
Deposition of Joseph Joice, who gave details of cannibalism that occurred in that county.
• Lack of consistency and predictability in EVERYTHING
  • Case
  • Punctuation
  • Spelling
  • Syntax
  • Semantics

• Domain knowledge resides with 17th Century Historians
  • Not Computer Scientists or Computational Linguists

• How to capture both structured, semi-structured and unstructured data?
<person>
  <forename>William</forename>
  <surname>Graham</surname>
  <type>Deponent</type>
  <town>Lisnemallet</town>
  <barony>Clankelly</barony>
  <county>Fermanagh</county>
</person>

<person>
  <forename>Ambrose</forename>
  <surname>Carleton</surname>
</person>

<commissioner>
  <name>William Aldridge</name>
</commissioner>

<general>
  <depositionID>835115nv008</depositionID>
  <date>8-Jan-1641</date>
  <manuscript>MS835</manuscript>
  <folio>f115</folio>
  <type>Dublin Original</type>
  <nature>Stripping</nature>
  <comments>very poor spelling!</comments>
</general>
LANGUAGE AND LINGUISTIC EVIDENCE IN THE 1641 DEPOSITIONS

...create a novel computer environment in which linguistic researchers can conduct sophisticated discovery, analysis and visualisation of the digitised 1641 Depositions.
• Apply LanguageWare, VARD, ManyEyes and Voyeur Tools

• New Insights

• Answers to ‘unanswerable’ questions

• Objective evidence to support subjective impressions
CULTURA

... pioneer the development and fusion of dynamic personalisation, contextual adaptivity and social analysis in a digital humanities context
NORMALISATION

\[ \text{traitorouslie} \]

\[ \text{traitorouslie} \]

\[ \text{traiterou} \quad \text{slie} \]

\[ \text{trai} \quad \text{ter} \]

\[ \text{ai} \quad \text{er} \]

\[ \text{tr} \quad \text{ai} \quad \text{er} \quad \text{ou} \quad \text{lie} \quad \text{ie} \]

\[ \text{trai:0.95} \quad \text{ai:0.8} \quad \text{er:0.7} \quad \text{ou:0.8} \quad \text{lie:0.85} \quad \text{ic:0.75} \]

\[ \text{0:0.75} \quad \text{1:0.7} \quad \text{2:0.65} \quad \text{3:0.55} \quad \text{4:0.75} \quad \text{5:0.5} \quad \text{6:0.7} \quad \text{7:0.6} \quad \text{8:0.65} \quad \text{9:0.8} \quad \text{10:0.7} \quad \text{11:0.55} \quad \text{12:0.5} \quad \text{13:} \]

\[ \langle \text{traitorously 0.8, traitorous 0.05} \ldots \rangle \]
ENTITY EXTRACTION
PERSONALISED RESEARCH ENVIRONMENT

For a continuum of end user types
CHALLENGE: CONSISTENCY

- Inconsistent ‘dirty data’
- Normalisation
- Harmonisation
- What do we do when there are gaps in the sources?
CHALLENGE: AUTHORITY

- Who decides what is correct?
- Mapping
- Merging
- Conflicts
CHALLENGE: REPRESENTATION

• Appropriateness for end user
• Standards (dis)agreement
• What is appropriate for what type of data?
• Mapping between and linking collections
• Linked Data
THANK YOU

Owen.Conlan@scss.tcd.ie and Seamus.Lawless@scss.tcd.ie

CNGL, School of Computer Science and Statistics, Trinity College Dublin

cultura-project.eu
cultura.fmi.uni-sofia.bg