the quality in quantity - enhancing text-based research

Bernie Ács, National Center for Supercomputing Applications, UIUC, USA
Andreas Aschenbrenner, State and University Library Goettingen, Germany
Tobias Blanke, Centre for e-Research, King's College London, UK
Patrick Harms, State and University Library Goettingen, Germany
Mark Hedges, Centre for e-Research, King's College London, UK
Felix Lohmeier, State and University Library Goettingen, Germany
Wolfgang Pempe, State and University Library Goettingen, Germany
Angus Roberts, University of Sheffield, UK
Kathleen Smith, State and University Library Goettingen, Germany
digital humanities

computer-aided
decades

shared infrastructure
years

e-Humanities

shared methodologies

collaborative forum
quantitative
comparative [breadth]
• (statistical) evaluation
• information extraction
• re-representation / visualisation

qualitative
source as such [depth]
• observing
• analyzing, understanding
• annotating

complimentary
TextGrid Services and Tools

XML-Editor
Graphical Link Editor
Workflow Editor
Search Tool
Dictionary Search Tool
Collationer
User and Project Management

Metadata Annotator
Streaming Editor
Lemmatizer
Text Publisher Web
Project Browser/Navigator
Tokenizer
Sort Tool
SEASR / MONK

SEASR (Software Environment for the Advancement of Scholarly Research)
MONK (Metadata Offer New Knowledge)
Andrew W. Mellon Foundation
Dunning Loglikelihood

• Feature comparison of tokens
• Specify an analysis document/collection
• Specify a reference document/collection
• Perform Statistics comparison using Dunning Loglikelihood

Example showing over-represented
Analysis Set: The Project Gutenberg EBook of A Tale of Two Cities, by Charles Dickens
Reference Set: The Project Gutenberg EBook of Great Expectations, by Charles Dickens
Text Clustering

• Clustering of Text by token counts
• Various filtering options for stop words, Part of Speech
• Dendogram Visualization
“The discussion of the children introduces each of the short internal narratives. This champions the view that her method of repetition was patterned: controlled, intended, and a measured means to an end. It would have been impossible to discern through traditional reading.”
Pattern identification using automated learning

- Which patterns are characteristic of the English language?
- Which patterns are characteristic of a particular author, work, topic, or time?
- Which patterns based on words, phrases, sentences, etc. can be extracted from literary bodies?
- Which patterns are identified based on grammar or plot constructs?
- When are correlated patterns meaningful?
- Can they be categorized based on specific criteria?
- Can an author’s intent be identified given an extracted pattern?
Dunning Loglikelihood Tag Cloud

- Words that are under-represented in writings by Victorian women as compared to Victorian men.
- Results are loaded into Wordle for the tag cloud
- —Sara Steger
why link qualitative and quantitative? they always have been linked ...

• create (one) - validate (many) research hypothesis (extrapolate)
• create (many) - validate (one) research hypothesis (replicate, show trends)
• explain / illustrate a trend (many) through individual examples (one)
• analyze an observation (one) through statistical analyses (many)
research lifecycle

inspired by
research lifecycle

- discover
- integrate
- re-represent
- analyze
- collate
- visualize
- drill-down
- contextualize
- explore

inspired by
finally

• challenges:
  1. get the data (automatic harvest or manual selection/upload?)
  2. integrate/normalise the data (semi-automatic?)
  3. get the analysis/visualisation right, along which dimensions?
• cue for the architecture:
  data will be redundant, to reuse existing systems and be open:
  (a) active use, (b) various analysis frameworks, (c) preservation
• usability: hide complexity!
  immediate results (automatic), and allow refinement (user)