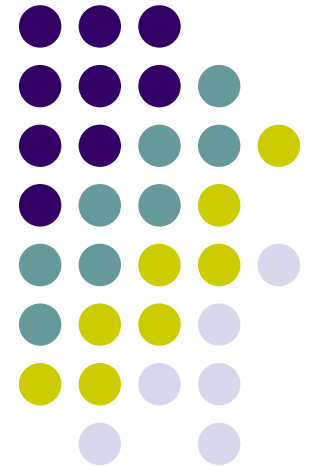


Reflections on e-science and linguistics

Paul Rayson
UCREL, Computing Department,
Lancaster University, UK.





Direct or indirect benefits?

- Could the main benefits to linguistics from e-research come indirectly via computational linguistics and NLP? E.g. via text mining?



Conferences and Workshops



- Text Mining, e-Research and Grid-enabled Language Technology
 - Fourth UK e-Science Programme All Hands Meeting (AHM2005)
 - September 2005, Nottingham, UK
- Towards a Research Infrastructure for Language Resources
 - Workshop at LREC, May 2006, Genoa, Italy
- NCeSS 2nd International Conference on e-Social Science
 - June 2006, Manchester, UK
- Historical Text Mining workshop
 - Supported by AHRC ICT Methods Network
 - July 2006, Lancaster, UK



Data Grid

- More, and more, and more data ...
- Most familiar to linguists
 - OTA, ELRA, LDC
- Framework of language resources for eHumanities
- LREC workshop on research infrastructure (2006)
 - standardization work: TEI, EAGLES, ISLE, MILE, ISO TC37/SC4
 - metadata frameworks: DC, IMDI, OLAC, MPEG7, METS
 - schemas: LMF, TIPSTER, EAF, MAF
 - knowledge representation: ISO DCR, GOLD
 - registration, integration and services: INTERA, TELRI, ECHO, DAM-LR, LIRICS
- CLARIN initiative (Common Language Resources and Technology Infrastructure)
- DAM-LR (Distributed Access Management for Language Resources)



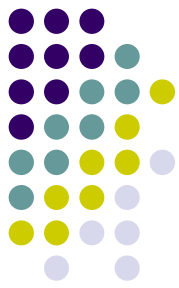
Access Grid

- Not just video conferencing ...
- ... large collaborative meetings
- Plug and play?
- NCeSS seminar series



Computational Grid

- Computing statistical language models from billions of words of natural language data
- Avoiding the corpus annotation bottleneck
- Workflow and configuration issues for NLP architectures



Semantic web

- Calzolari (2006) “cooperation must be enhanced among many communities acting now separately, such as LR and LT developers, terminology, SW and ontology experts, content providers, linguists, humanists.”
- Semantic mark-up
- Ontology extraction
- Language resources as web services
- Open access
- “Spinning” the semantic web



Case study

- Changing English Across the Twentieth Century: a corpus-based study
 - <http://ucrel.lancs.ac.uk/20thCenturyEnglish/>
 - Sponsored by The Leverhulme Trust (Grant number F/00 185/J), this project runs from August 2005 - July 2007.
- Standard written British English
- Sampling dates 1901 – 1931 – 1961 – 1991
- Each corpus contains 1M words
- 15 genres of published informative and imaginative prose (e.g. press reportage, academic writing, romantic fiction, science fiction)
- Corpus size and sampling frame modelled on the Brown Corpus (of 1960s American English)

Comparable corpora across the C20th

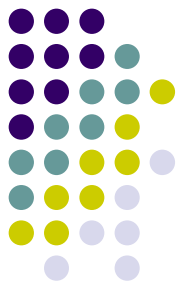


	1901 (1898-1904)	1931 (1928-34)	1961	1991/ 1992
BrE	Lanc-1901	B-LOB (Lanc-1931)	LOB	F-LOB
AmE	?	Pre-Brown31	Brown	Frown



Project stages

- Data collection
- Conversion
- Encoding
- Annotation
- Indexing
- Analysis

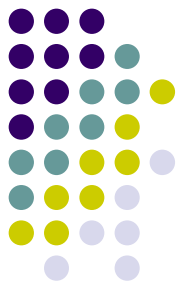




Data collection

- Following sampling frame of LOB/Brown family
- Selection (BL, DSC, Collindale and Manchester newspaper libraries, Gutenberg, Times Archive)
- Check author is British
- Copyright clearance

Conversion and preparation



- Scanning and OCR
- Copy typing
- Encoding and annotation
- Indexing in IMS-CWB
- Concordancing and manual classification

Analysis



- Key concepts in 20th C. romantic fiction
 - Bottom up
- ‘Obligation & necessity’ in 20th C. corpora
 - Top down



Grand challenges

- If a linguist wants to count obligation and necessity terms in the BNC and LOB, then he/she needs to use several different tools, know several tagsets and a specific regular expression formalism
- Larger corpora (size)
- Historical corpora (variants)
- When this analysis has been done once, can it be made available?
- Google-like ease of use for tools?
- Enabling the conversations: an e-linguistics forum