

AHRC e-Science Scoping Study Final report: Findings of the Expert Seminar for Linguistics

Paul Rayson
Director of UCREL
Lancaster University, UK
14th August 2006

1. Introduction

This document summarises the discussions that took place and the conclusions drawn during the e-Science Scoping Study Expert Seminar for Linguistics. The Linguistics Expert Seminar took place on one of the hottest days of the year at the Arts and Humanities Data Service, Drury Lane, London on Tuesday 18th July 2006. The participants were as follows:

- Svenja Adolphs (University of Nottingham)
- Sheila Anderson (AHDS)
- Tobias Blanke (AHeSSC)
- Luke Blaxill (AHDS)
- Lou Burnard (University of Oxford) [Unable to attend but contributed to the discussion via email]
- Mark Davies (Brigham Young University)
- Willem Hollmann (Lancaster University)
- Christian Kay (University of Glasgow)
- David Nathan (SOAS)
- Wim Peters (University of Sheffield)
- Paul Rayson (Lancaster University) [Chair]
- Ann Taylor (University of York)
- Martin Wynne (AHDS Literature, Languages and Linguistics)

The seminar began with an introduction by Sheila Anderson, followed by a discussion framed by a document circulated in advance of the seminar and written by Paul Rayson entitled “A personal take on the ICT challenges and barriers in Linguistics”¹. After lunch, Luke Blaxill gave an introductory presentation on e-Science followed by case study presentations by Tobias Blanke and Svenja Adolphs. Paul Rayson presented his reflections on e-Science and Linguistics². In the discussion that followed, the participants contributed to a wish list of projects, tools and initiatives intended to address the key challenges identified earlier in the day and contributing to an e-Arts and Humanities agenda. The seminar was concluded with a welcome glass (or two) of wine.

¹ See <http://www.ahds.ac.uk/e-science/documents/Linguistics-grand-challenges.pdf>

² See <http://www.ahds.ac.uk/e-science/documents/Rayson-presentation.pdf>

2. The eSSS project

The context for the e-Science Scoping Study (eSSS) project³ is the Arts and Humanities Research Council (AHRC) funded ICT (Information and Communication Technology) programme managed by David Robey at the University of Reading. Part of the ICT programme has been to develop the e-Science agenda for the Arts and Humanities. Twelve ICT strategy projects have been funded along with the AHRC ICT Methods Network. Other funding agencies such as EPSRC and JISC together with AHRC have put resources into the e-Science research projects scheme and there will be an open meeting in London on 8th September 2006 to provide an opportunity to ask questions about the scheme. Six four-year PhD project studentships are also being made available. The eSSS project itself came about via a seminar in April 2004 chaired by Sheila Anderson in the area of e-Science in the Arts and Humanities. The aims of the eSSS project are to:

1. Raise awareness and understanding of e-Science and to work with smaller groups of experts to define what we mean by e-Science
2. Raise the profile of e-Science and the use of advanced ICT methods in the Arts and Humanities
3. Assist AHRC in its development of the e-Science agenda for the Arts and Humanities by identifying priority areas for research and practise

3. Digital Linguistics

Information and communications technology (ICT) and, in particular, computing technology has already had a huge impact on linguistics. Empirical linguistics experienced a revival in the late 1970s and early 1980s and this coincided with wider availability of personal computers, corpus software, the creation of electronic corpora and the increased use of corpus-based techniques. The impact of ICT is still ongoing alongside new developments in ICT itself, such as processing power, data storage, and network connectivity.

In order to discuss whether “Digital Linguistics” already exists as a discipline, the seminar chair placed the various challenges and barriers for extending the use of computing technologies in the context of the corpus linguistic research process. McEnery and Wilson (1996) defined corpus linguistics as a methodology that can be applied to a wide range of linguistic study. This is ably demonstrated by Biber, Conrad and Reppen (1998) who described corpus-based approaches in different areas of linguistics, including lexicography, grammar, discourse, register variation, language acquisition and historical linguistics. Paul proposed a model in which there are five core steps in the corpus-based approach:

1. **Question:** A research question or model is devised
2. **Build:** Corpus design and compilation
3. **Annotate:** Computational analysis of the corpus
4. **Retrieve:** Quantitative and qualitative analyses of the corpus

³ See <http://www.ahds.ac.uk/e-science/e-science-scoping-study.htm>

5. **Interpret:** Manual interpretation of the results or confirmation of the accuracy of the model

There are at least three further stages to the research process that are typical across many, if not all, disciplines:

6. **Output:** Distil the research into a paper or presentation
7. **Dissemination:** Pass the paper to a publisher for printing, or submit to a conference for presentation
8. **Feedback:** Reviews of papers or presentations, citation practice influence the direction of future research

As outlined above in these five core steps, the process model of the corpus-based methodology typically used by researchers is as follows: it would begin with the identification of a research question (1), continue with building (2) and annotating (3) a corpus with which to investigate the topic, and finish with the retrieval, extraction (4) and interpretation (5) of information from the corpus which may help the researcher to answer the research question or confirm the parameters of the model. In some cases, the process may be an iterative one, where, following the interpretation of the results some refinement is needed on the research question or annotation of the corpus.

During the following discussion, the following points emerged:

1. ICT has already had a major impact on linguistics and on corpus linguistics in particular. ICT is used in every stage of the process model
2. The discussion ranged around how good the proposed process model was and how suitable it is if we are using usage-based models or a theoretical linguistics paradigm
3. The annotation phase may be omitted completely
4. The annotation phase may proceed along manual lines rather than computational analysis via automatic tagging software
5. The level of annotation depends on the research question
6. Other sub-disciplines of linguistics may have different process models, e.g. in documentary linguistics (language documentation), the corpus is the output of the research process rather than the input
7. In some sub-disciplines, there may be no accepted process model and research progresses in a way defined by the individual scholar
8. The above process model was developed on the assumption that the language has been transcribed and does not include audio and video recordings of language. In the British National Corpus spoken sub-corpus, for example, sound archives do exist but are not readily available. However, recent developments are changing this assumption. In the SCOTS project⁴, audio and video data are now linked directly to concordance lines. Multi-media corpus linguistics poses great technical challenges but also great opportunities to ask new research questions that haven't been possible before, e.g. investigating the storage of multiword expressions (MWE) by native

⁴ <http://www.scottishcorpus.ac.uk/>

and non-native speakers by examining the placement of pauses near the MWEs in speech (Second Language Speech Fluency and Multi-word Units Project⁵)

9. Increasingly, researchers do not need to build and annotate (steps 2 and 3 above) their own corpus material. Instead they can use precompiled and annotated corpora that are available 'off-the-shelf' (Meyer 1991). Data providers and corpus builders undertake steps 2 and 3 and the corpus users undertake steps 1, 4 and 5. This results in a disconnection between the data providers and the corpus users. The seminar group also found that this is mirrored on a wider scale with content providers such as Proquest, Google and the British Library not understanding (or being aware of) the requirements of linguists
10. There is a missing step: depositing the corpus in an archive for reuse
11. Theory should sit alongside each step and should back up the interpretation stage. Other methods of data collection, for example in syntax, are elicitation experiments. ICT is used in the creation of typological databases using spreadsheets
12. A key challenge would be to bring linguists and computer scientists (CS) together and involve the CS people in research that will be recognised in their own discipline as much as it would be for the linguists, in the context of the RAE for example

4. ICT challenges and barriers in Linguistics

Since the first electronic corpora were collected in the 1960s and 1970s, the size of the largest corpora has roughly increased by a factor of 10 every decade. We can see this trend from the Brown and LOB corpora (1 million words in the 1960s and 1970s), through the British National Corpus (100 million in the 1990s), to the Oxford English Corpus (1 billion words in 2006). A key challenge is making this amount of data usable for individual scholars sitting at their desktops, and facilitating such collection activities by those scholars. The use of the web as a corpus is becoming more common place⁶. How can such large amounts of data be cleaned, encoded, annotated, stored, and shared are all key questions. Clearance of copyright for web data as well as other corpus data is a vital issue. Challenges for annotation are related to problems of tagsets and scale.

The following points emerged from the discussion:

1. How acceptable is it to reuse standard annotation schemes on new data?
2. How can we avoid the annotation bottleneck for the web as corpus (Rayson et al, 2006)?
3. Can we undertake collaborative manual annotation and share the effort and results via the internet?
4. New architectures such as relational databases are needed to deal with the large scale corpus resources

⁵ <http://www.nottingham.ac.uk/english/research/cral/adolphs.htm>

⁶ http://sslmit.unibo.it/~baroni/web_as_corpus_eacl06.html

5. In relation to copyright, better promotion of good codes of conduct, more awareness and recognition of what is possible within the law (e.g. by adoption of open source licensing policies) is required

Other disciplines such as digital humanities (humanities computing, literary studies) have their own well known tools (such as TACT) that offer similar functionality (Hockey, 2000). In the area of Computer-Assisted Qualitative Data Analysis Software (CAQDAS) similar tools are available for content analysis, discourse analysis and frame analysis methodologies⁷. Social science researchers have yet another view of the tools available for text statistics and concordancing⁸. We can distinguish three different generations of text retrieval software:

1. **Dos and Command-line:** WordCruncher, OCP, Childes-Clan, LEXA, TACT, MicroConcord
2. **Graphical user interfaces:** IMS-CorpusWorkBench, ICECUP, Xaira, WordSmith, MonoConc
3. **Web-based:** VIEW, BNCweb, SketchEngine, Wmatrix, PIE

The following points were covered during the discussion:

1. There has also been a shift from black-box software packages to more service-oriented architectures such as in Xaira and eXist which enable developers to construct their own interfaces (both machine-machine and human-machine) and rely on open standards for data communication
2. Will the next generation of retrieval software be grid based?
3. Can we open a dialogue with these other disciplines to produce better applications suitable for multiple users, or will general purpose tools be of any use for a specific purpose? Do you first need to decide on which methodology you want to use to analyse your data, and then match your analysis with the appropriate software?
4. Web-interfaces may restrict access to the full data, if someone needs it for their processing but we may require new ways of working where you define your computational processing and have it done remotely via the grid
5. Copyright restrictions on access to the full text: the BNC consortium didn't consider obtaining permission for web distribution because the web didn't exist when the BNC project first started
6. Data providers need to get permission to be re-contacted, in order to ask for permission for new types of research or new types of distribution.
7. Empowering individuals to be data providers, this is already happening for example in projects like BBC World War Two memories and Voices.

⁷ <http://www.lboro.ac.uk/research/mmethods/research/software/caqdas.html>

⁸ <http://www.lboro.ac.uk/research/mmethods/research/software/stats.html>

For a wider review of text analysis software see <http://www.textanalysis.info/>

5. e-Science definitions

e-Science is something of a buzzword, and is not easy to define. For some, it's simply another name for grid technologies. For others, the term covers most of computer assisted research, hence the popular, and much broader, alternative name 'e-research' which has achieved fairly widespread adoption in the social sciences.

In the forthcoming AHRC Glossary for ICT, however, e-Science has been defined as-

"developing shared access to research facilities distributed across the Internet in the form of computational processing and data collections. This has allowed more powerful and innovative research designs in many areas of scientific research, and is capable of making substantial differences the arts and humanities as well"

This flexible definition recognises that it is as much a methodology as a system or toolkit. At its most basic level e-science is about sharing resources in a secure networked environment. These resources might be computational processing power, expertise, manpower (or womanpower), or data. This is different from the internet as it is now, which is designed more for sharing small documents than any of the aforementioned resources. Naturally, the power to electronically share such resources from anywhere on the globe is so considerable that e-science, if it could fulfil its potential, might revolutionise the way we work in many areas of the arts and humanities in the future.

In his presentation, Luke stressed that e-science is a quite a new idea, and is not particularly evolved even in the hard sciences. Most of the actual e-science developments to date, however, fall under the heading of grid technologies. The term is borrowed from the national electricity grid, a macro infrastructure that delivers scalable amounts of electric power to consumers from the larger commercial user to the smallest domestic household. It doesn't matter how or where the power is generated; you flick a switch, and you get power anywhere accessible by the National Grid. Computing Power is similar, except that the e-Science grid is global and is delivered via the fibre-optic cable infrastructure of the internet. There are 3 forms of e-science grid:

1. The access grid: likely to be the most familiar form of grid to Arts and Humanities researchers. It enables large-scale meetings between geographically distributed researchers via video conferencing.
2. The data grid: concerned with linking of existing resources. The data grid is perhaps the most relevant to the Arts and Humanities
3. The computational grid: concerned with High Performance Computing. The application of this technology in the Arts and Humanities is more limited and concerned with high performance visualisation of very large datasets

Following these definitions, the group saw some sample e-science projects and tools from the forthcoming eSSS Knowledge Base. TextGrid⁹, for example, TextGrid aims to create a community grid for the collaborative editing, annotation, analysis and publication of specialist texts. Svenja Adolphs presented the Digital Record project¹⁰. Digital Record is a node of the National Centre for e-Social Science based at Nottingham. The project involves social scientists and computer scientists on three driver projects to develop e-social science applications demonstrating the salience of new forms of digital record.

6. Reflections on e-Science

Paul presented his reflections on how e-Science might be harnessed to meet the challenges in Linguistics identified during the seminar. He speculated that the benefits to linguistics might come indirectly via the computational linguistics and natural language processing (NLP) communities, for example via the research into text mining. Conferences and workshops such as the following do enable some cross over between the communities, but more events such as these are required:

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

Calzolari (2006) highlighted this need for communication between different research communities: “cooperation must be enhanced among many communities acting now separately, such as LR [Language Resources] and LT [Language Technology] developers, terminology, SW [Semantic Web] and ontology experts, content providers, linguists, humanists.”

In terms of the opportunities presented by the grid itself:

1. The data grid: most familiar to linguists are organisations such as the Oxford Text Archive¹¹, European Language Resources Association¹², and the Linguistic Data Consortium¹³. How do resources such as these become grid enabled? Does this enforce a common standard and

⁹ <http://www.textgrid.de/>

¹⁰ <http://www.ncess.ac.uk/nodes/digitalrecord/>

¹¹ <http://ota.ahds.ac.uk/>

¹² <http://www.elra.info/>

¹³ <http://www ldc.upenn.edu/>

- interface for all corpora? Would this preclude development of techniques to do something new with a specific resource?
2. The access grid: large scale meetings via video conferencing facilities in Access Grid Node may be the first contact that linguists have with the grid, for example in the NCeSS seminar series. Ease of use for these facilities is important to ensure that the technology creates a bridge rather than a barrier.
 3. The computational grid: computational linguistics researchers are already working with highly data intensive methods and these are prime candidates for grid computing. Combined with billions of words of natural language data, this facilitates the computation of large statistical language models. There are still problems to be solved with the corpus annotation bottleneck and workflow/configuration issues for NLP architectures

The Leverhulme funded project “Changing English Across the Twentieth Century: a corpus-based study”¹⁴ underway at Lancaster University was presented as a case study to see how e-Science methods might change the research methodology itself. The project involves extending the existing pair of 1-million word corpora LOB (1961 British English) and FLOB (1991 British English) to two new times frames, 1901 and 1931. Paul speculated that in the grid enabled future, we might be able to employ the data grid to build a large number of equivalent corpora. It might be possible to bypass the existing collection method which involves trips to the British Library, Collindale and Manchester newspaper libraries to select and copy each one of the 500 text samples, before scanning or copy typing by hand. Multiple versions of each corpus could be annotated, compared and analysed in parallel using the computational grid. This development might in fact turn on its head the notion of a static representative corpus.

7. e-Linguistics agenda

The following items were highlighted by the seminar participants as important research needs which might feed into the e-science agenda:

- Data acquisition and management:
 - A requirement for training linguists in data management, e.g. standards for data archiving, curation or sharing
 - Provision of a service for data digitisation and encoding to sit alongside the existing archives e.g. LDC, OTA, ELRA
 - Open source corpus development to allow collaborative effort to digitise, encode or annotate linguistic data or corpora
 - Software to perform format conversion (encoding) or tagset mapping (annotation)
- Data annotation
 - Collaborative annotation tools to allow multiple researchers to annotate the same text in order to compare and comment on all the interpretations, e.g. metaphors

¹⁴ <http://ucrel.lancs.ac.uk/20thCenturyEnglish/>

- Could we exploit the web service pricing model to achieve this?
- Data access
 - Communication with content providers such as Chadwyck-Healey, Proquest, Google, British Library to encourage an understanding of what our requirements are for accessing their data, e.g. within EEBO, LION and Times Archive
 - Communication with JISC who may be able to negotiate on behalf of the linguistics community to licence data
 - Access to large digitised resources in a usable format e.g. EEBO, Times Archive, 19th century books
- Data retrieval
 - Software that is easy to use “Google-like” (for all corpora and the OED)
 - Multimedia (direct access to the audio and video rather than the transcribed version). What new kinds of research might this open up?
 - “Web as Corpus”: dealing with problems of size
 - Historical corpora: dealing with variant spellings

8. Conclusion and e-science opportunities

Corpus linguistics has in the past focussed on English and in particular modern standard varieties of English. Challenges that have started to be addressed in the last five to ten years are the adaptation of the techniques and tools to non-English, historical and dialectal corpora (Archer et al, 2003; Beal et al, 2006). Utilising the results of the vast amount of digitisation activity that is being undertaken by commercial organisations: (e.g. Open Content Alliance, Google Print, Early English Books Online) is one of the biggest challenges facing corpus linguistics over the next few years. Methods such as frequency profiling, concordancing, n-grams and keywords will need to be made scalable. Problems of unreliability may surface due to, for example, variation in spelling, when these techniques are applied to the voluminous amounts of historical and other corpora that will be at our disposal in the near future. If linguistics as a discipline is to take advantage of these new resources, e-Science may provide the tools and techniques to do so. Sharing the pockets of expertise gained from individual projects is of vital importance, and the eSSS Linguistics expert seminar provided one such forum. Linguistics is well placed to take advantage of the e-Science opportunities through the extension of advanced ICT methods and tools already taken up in the corpus and computational linguistics communities.

Acknowledgements

The Expert Seminar in Linguistics formed part of the E-Science Scoping Study project directed by Sheila Anderson and was organised by Luke Blaxill (Project Research Assistant) and Katrin Weidemann (Project Administrator). Presentations were given at the seminar by Sheila Anderson, Luke Blaxill, Tobias Blanke and Svenja Adolphs. Their assistance and the contributions of

the other seminar participants is gratefully acknowledged. Further details of the E-Science Scoping Study project can be found at:
<http://www.ahds.ac.uk/e-science/e-science-scoping-study.htm>

References

- Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In *Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University, pp. 22 - 31.
- Beal, J., Corrigan, K., Rayson, P. and Smith, N. (2006) Writing the Vernacular: Transcribing and Tagging the Newcastle Electronic Corpus of Tyneside English (NECTE). *Pre-conference workshop on corpus annotation, ICAME-27*, University of Helsinki, Finland, 24 May 2006.
- Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: investigating language structure and use*. Cambridge University Press, Cambridge.
- Calzolari, N. (2006). Community Culture in Language Resources – An International Perspective. 2. Towards a Research Infrastructure for Language Resources, Workshop in conjunction with LREC, May 2006, Genoa, Italy, pp. 12 – 15.
- Hockey, S. (2000). *Electronic texts in the humanities*. Oxford University Press.
- McEnery, T., and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh University Press, Edinburgh.
- Meyer, C. F. (1991). A corpus-based study of apposition in English. In Aijmer, K. and Altenberg, B. (eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. Longman, London, pp. 166 – 181.
- Rayson, P., Walkerdine, J., Fletcher, W.H. and Kilgarriff, A. (2006) Annotated Web as corpus. In proceedings of the 2nd Web as Corpus Workshop held in conjunction with the *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, April 3, 2006, pp. 27 - 33.