Panel Discussion on Computing and the Humanities

John Nerbonne, Paul Heggarty, Roeland van Hout and David Robey

This is the report of a panel discussion held in connection with the special session on computational methods in dialectology at *Methods XIII: Methods in Dialectology* on 5 August, 2008 at the University of Leeds. We scheduled this panel discussion in order to reflect on what the introduction of computational methods has meant to our subfield of linguistics, dialectology (in alternative divisions of linguistic subfields also known as variationist linguistics), and whether the dialectologists' experience is typical of such introductions in other humanities studies. Let's emphasize that we approach the question as working scientists and scholars in the humanities rather than as methodology experts or as historians or philosophers of science, i.e. we wished to reflect on how the introduction of computational methods has gone in our own field in order to conduct our own future research more effectively, or alternatively, to suggest to colleagues in neighbouring disciplines which aspects of computational studies have been successful, which have not been, and which might have been introduced more effectively. Since we explicitly wished to reflect not only on how things have gone in dialectology, but also to compare our experiences to others, we invited panellists with broad experience in linguistics and other fields.

We introduce the chair and panellists briefly.

> John Nerbonne chaired the panel discussion. He works on dialectology, but also on grammar, and on applications such as language learning and information extraction and information access. He works in Groningen, and is past president of the *Association for Computational Linguistics* (2002).
>
> David Robey is an expert on medieval Italian literature, professor at Reading, and Programme Director of the *ICT in Arts and Humanities* Research Programme of the (British) Arts and Humanities Research Council. He is also president of the *Association for Literary and Linguistic Computing*.
>
> Paul Heggarty is an expert on the Quechua language family, and also works on Romance and Germanic, studying both synchronic variation and diachronic developments, where he's one of a growing number of researchers applying phylogenetic inference (and software developed in biology for this purpose) to historical linguistics. He's at Cambridge.
>
> Roeland van Hout is an expert on sociolinguistics, phonetics and second language acquisition. He is co-author (with Toni Rietveld) of the world's best book on applying statistics to linguistics, (Harald Baayen has just come out (2008) with a contender, but 15 years at #1 is impressive). He's at Nijmegen, and he's served on the Dutch research council's Humanities board for the last five years.

We suggested in inviting the panellists that they deliver some introductory remarks structured in a way that responds *inter alia* to two questions:

> Are there elements of the computational work in language variation that should be made known to other scholars in the humanities?
>
> Are there other computational lines of work in the humanities that are not used in studying language variation but which should be of interest? Is the primary common interest the computational tools or are there research issues which might benefit from a plurality of perspectives?

We emphasized, however, that the panel speakers should feel free to respond to the question of how the disciplines might interact in more creative ways, especially if they feel that the questions above are put too narrowly.

Each panellist presented some ideas before the discussion started.

***Paul Heggarty** summarised some basic methodological issues in applying computational techniques to language data, reviewing typical criticisms — and how best to answer them.*

A burgeoning interest in the use of computational techniques is certainly something that linguistics holds in common with other fields in the humanities. But what might the various disciplines learn from each other in this? Here I offer a perspective from the experience of attempts to apply such techniques to comparative and historical linguistics, to be taken here as an illustrative “humanities data-type”. Not that this experience has always been plain sailing, so some of the “lessons” turn out to be more in how *not* to go about things.

An important aside to begin with is to clarify that what the disciplines learn from each other is hardly limited just to the methods, but extends also to the *results* from applying them. Computational approaches can have a particularly powerful impact when they turn out to challenge traditional views which, while entrenched, have hitherto remained unquantified and essentially impressionistic. Such is the case with Gray and Atkinson’s (2003) article in *Nature*: while much disputed, to be sure, it has undoubtedly forced open again the question of the dating and origins of Indo-European, and a possible link with agriculture as a driving force. The relevance of this particular application of computational techniques to language data, then, transcends linguistics into several other branches of the humanities and beyond, not least archaeology, (pre-)history and genetics.

To focus on methods, though, we need to consider not only which tools to use, but more general issues of methodology, two in particular. Firstly, how freely transferable are computational methods between disciplines? And secondly, it is not just about which methods to use, but also what one does with them, how imaginative one has to be in order to make the best use of them.

A good illustration of transferability is provided by phylogenetic analysis algorithms for reconstructing ‘family trees’ of descent. While initially devised for applications in the life sciences, particularly speciation, genetics and population studies, these methods are now increasingly applied in historical linguistics and dialectology, while further examples from across the humanities include comparative studies of the histories of manuscripts, cultural artefacts and religions.

From these illustrations there might appear to be no limit to the potential uses of such methods across disciplines. Indeed, one widespread view is that computational methods are *per se* generic, and as such "discipline-neutral" and therefore applicable in principle to any field. Yet for other scholars the whole enterprise of applying computational methods in the humanities is undermined by too automatic an assumption of transferability. This school argues that one cannot just take existing computational methods "off the peg", for especially if they were originally designed for the data in the natural sciences, they simply may not be meaningfully applicable to the forms of data proper to the humanities, such as language variation data.

There is, however, a sense in which it is possible for both views to be valid, albeit at two quite different stages. To understand how this can be so, one must distinguish component stages within the process of applying computational techniques to data from another discipline, as follows.

Stage 1: *Encoding*: getting from "raw" real-world data to expressions of them in numerical or state data format, suitable for computational analysis.

Stage 2: *Processing*: further 'number crunching' on the now encoded data, typically statistical or phylogenetic analyses.

Stage 3: *Interpretation* of what the results from such processing mean, back in the real world: e.g. for the divergence histories of languages.

For our question of transferability, what matters is that by the end of stage 1, once the data have been converted into state or numerical format, then the link has been broken with whatever original 'real-world' form those data take in any particular discipline. So from then on, any further mathematical processing, of what are now effectively just numbers, is indeed intrinsically field-neutral. It follows that for stage 2 processing it generally is safe to take computational techniques "off-the-peg", even if originally developed for applications in other disciplines.

Not so for methods for the preceding stage, however, during which the original raw data — such as language data on sounds, vocabulary or grammatical structures — are first encoded into state or numerical format. For, in marked contrast to the natural sciences, social science data generally do *not* lend themselves to straightforward expression either in numerical terms or as a set of state values all equally different to each other. Long-standing objections to lexicostatistics, for instance, question whether relationships and entities in natural languages really are discrete enough for the method's "straightjacket" of all-or-nothing yes/no data-points to constitute valid encodings of a complex linguistic reality, rather than doing violence to it.

A poor stage 1 method, then, is open to the charge that it positively <u>mis</u>represents the linguistic reality. Indeed, much of the scepticism towards attempts to apply computational methods in linguistics boils down to a widespread perception that (to paraphrase the typical refrain) simply "bandying numbers about" is rather "too easy a game to play".

Representations need to be grounded from the outset in a sound and principled methodological approach to the whole enterprise of encoding. Particularly in a context where one needs to inspire greater confidence among sceptics, computational methods are best accompanied by a frank and explicit recognition that no particular encoding can claim necessarily to represent *the* definitive, absolute numerical expression of linguistic facts. To be realistic, we can aspire to no more than a best approximation, as meaningful a mathematical interpretation of linguistic concepts as can reasonably be achieved. As an illustration, take the task of measuring distance between languages on the level of sound: naturally, any solution must be "phonetics-led", with weightings based upon and reflecting the true relative significance of the various types of difference in phonetics, as per the principles and architecture of phonetic analysis. An off-the-peg computational method, by contrast, with no particular relation to the sounds of human language, is unlikely to convince sceptics that the numbers it churns out are valid approximations to linguistic reality.

Furthermore, it turns out that there are even some limits to the universal applicability of stage 2 processing methods too. One type of further processing, increasingly popular in recent years, is phylogenetic analysis. As applied in historical linguistics, at the simplest level this can be seen as attempting to reconstruct from data on the divergence of related languages the "family tree" by which they most plausibly "evolved" into the particular relationships now observed between them. Some of these very terms, however, are ones with which many sceptical linguists are uncomfortable, and it is here that the question of transferability can resurface, if in another form, even for stage 2 techniques. For each of the various phylogenetic analysis methods embodies some particular *model* of 'evolution' (of character states); but do these models necessarily match how the entities in any particular humanities discipline (e.g. languages, or their words and sounds) 'evolve' in reality? Most phylogenetic analysis packages were originally developed for applications in the biological sciences; but do speciation, genetic or population evolution represent realistic models also for the change and divergence processes in any given humanities discipline?

To close, I venture a few tips that might serve other disciplines, drawn from the experience of attempts to apply computational techniques to language, and the typical responses from sceptical critics.

Response 1: "Publish and be damned!". Certainly, computational and especially phylogenetic analyses appear to offer a rare opportunity for work in humanities disciplines such as historical linguistics to break into the leading world journals for the natural sciences, such as *Nature*, *Science* or *PNAS*. Yet while this may be a way to acquire a reputation, it by no means guarantees that it will be a glowing one within the humanities discipline in question. Linguists, for instance, have frequently taken a dim view of such approaches to language data — often applied by non-linguists, and "computation-led" rather than linguistics-led.

Response 2: "Lies, damned lies, and statistics". The objection here is that it is all too easy to churn out *some* numbers, and then to make claims for what they "demonstrate", when in fact the numbers can equally well be interpreted in quite different ways. Witness for instance Nichols' (1990) claim of linguistic support for an *early* date for the population of the Americas, to which Nettle (1999) ripostes that Nichols' own figures are in fact just as compatible with a *late* date, and really say nothing either way. One must clearly be honest and realistic in one's claims, and mindful of how heavily they may rest on any debatable assumptions. Another cautionary tale is that of glottochronology, widely discredited for its base assumption of a 'clockwork' rate of change over time, observable in practice to be quite false.

Response 3: “So what?” This final form of reaction is reflected in how, as already observed, sceptics frequently object that the numbers and graphics produced in computational studies are, for specialists in that field, neither useful nor informative, and represent little more than another way of portraying what we already know, without saying anything new.

How can one go about countering all these objections, particularly the third, to ‘sell’ the utility of computational techniques to a sceptical discipline?

One approach is to target as directly as possible particular research questions as yet unresolved, especially where debate has continued to rage between unquantified, impressionistic and thus potentially subjective positions. It is particularly in such cases that computational methods can be shown to offer a valuable new perspective and objectivity. Gray and Atkinson (2003), for instance, squarely addressed the age-old problem of the time-depth of the Indo-European language dispersal, recruiting for this specific purpose an approach to dating never previously applied to the question. Or, closer to home for the purposes of this issue, Maguire *et al.*’s paper in this volume devises statistical tests of subset diversity to investigate whether two particular topical claims in English dialectology are actually borne out in reality: is ‘dialect levelling’ really going on, and is there such a thing as a coherent ‘Northern Englishes’ dialect grouping?

Where quantitative and further processing methods really come into their own, offering undeniable advantages over impressionistic and unquantified judgements, is in handling great masses of data. Large datasets on variation typically encompass multiple complex and conflicting signals which cannot all be weighed up against each other simply by “eyeballing” them; indeed attempting this usually tempts us towards unduly simple analyses easier to ‘get our heads around’.

In any case, to convince sceptics of the utility of computational methods in the humanities, it is not just a question of *which* method(s) to use. Rather, much hangs also on precisely *how* they are put to use: it pays to be as imaginative as possible in order to get the very most out of all these techniques. Above all, a key alternative approach for cutting through masses of complex variation data is not to look only at overall results for a database as a whole, but to home in on particular research questions by isolating multiple independent sub-sets of data within it, and then comparing and contrasting them.

Many such inventive applications are pursued in this volume, which together attest to the growing enthusiasm for computational techniques, both in dialectology and in linguistics more generally. It is hoped that their successes and teething troubles alike, as reviewed here, might serve to help other disciplines to harness the undoubted potential of such methods — when carefully applied — far more widely across the humanities.

Roeland van Hout *calls for more reflection on the choice of computational methods and more attention to work in closely related fields.*

It is very sensible to consider the relation of computational techniques in the study of language variation to their use in other humanities disciplines although it is only a small field compared to humanities in general. Nevertheless, the application of all kinds of computational tools is common in the area of language variation research. What can we learn from each other as we continue to expand the remit of computational work in the humanities?

It should be clear that the role of the computer is not only central to the study of language variation but also that it has expanded enormously in different parts of the humanities over the past five or ten years. The computer provides access to data, which we see reflected in the much larger datasets that are now routinely analysed, and we have seen good examples of the benefits of broader access at this workshop. The computer most obviously improves the opportunities for sophisticated access, data analysis and visualization, especially visualization based on statistical analysis, again, an improvement we have benefited from in the past two days. In the earlier days of dialect geography straightforward symbol maps or maps with isoglosses were the main visualisation instrument. Nowadays, we have all sorts of maps and graphical representations, and we know better how different data sources can be combined, recombined and integrated in language variation studies. The opportunities to open up, combine and integrate various rich data sources (e.g. historical, geographical, social, political, linguistic), again and again, opens new vistas for doing research in the humanities. It will be a main trigger in developing e-humanities.

Finally we suspect that simulation will play an increasing and decisive role in scholarly thinking and argumentation, including the field of language variation, and simulation is naturally completely dependent on computers. Computer simulations are more than a play, they give the researcher in humanities the possibility to develop and test theoretical claims, as we can observe in many other scientific fields nowadays, in different disciplines. There is no reason to assume that this way of scholarly thinking would not pervade the humanities. I note all of this as a background to a criticism, and, in particular to emphasize that my criticism appreciates the enormous advantages computational techniques have brought to the study of language variation.

Let me take the opportunity to note a couple of points in the study of language variation and its use of computers where I think more progress might be expected and where therefore more attention might be paid. First, many of the computational techniques probe language and dialect differences at a fairly high level of aggregation, while most linguistic theory concerns smaller units of language, say, words, sounds, features and constraints. So it looks as if there is an opportunity to pay more attention to the role of individual items of variation in these larger scale analyses. This might be important as we attempt to link the study of language variation to the study of variation in other aspects of culture. Second, we see different lines of research using different sorts of data, some proceeding from abstract, theoretically motivated linguistic features and others proceeding from more concrete, intuitively accessible material. Naturally this stimulates the question of how these relate to each other. Third, we have also seen linguistic variety approached from a large number of perspectives using a variety of measures, including acoustic measures, pronunciation distance using Levenshtein distance, lexical heterogeneity, and even measures involving the informed opinions of lay dialect speakers. What intrigues me about this variety linguistically is the varied inroads that might open to understand linguistic structure and linguistic systems. Fourth, let me segue to more general issues by noting that even while we acknowledge the multi-dimensionality of linguistic variation, many successful approaches to its study focus on the lexicon where these many dimensions are reflected. This leads me to speculate that lexical studies may also serve as the most important bridge to other disciplines, as well.

I wish to turn then to the more general issues of how we might learn from each other in the different humanities disciplines. When I consider how we might present the results of our work to fellow scholars in the humanities, then several remarks are appropriate. First, we need to distinguish the development of a measure of difference from an account of all that a difference might entail. Levenshtein distance appears to be an effective technique for getting at pronunciation differences, but this in not the same as an account of the production or perception of these differences in the individual. Perhaps the intelligence test might be an appropriate analogy. The intelligence test measures intelligence, most agree, but that doesn't mean that it sheds light on the cognitive operations required for intelligent behaviour. I think it is useful to keep this in mind as we present our work to neighbouring disciplines.

Second, it is my impression that the use of computational techniques in the study of language variation is still young and exploratory, and that a period of reflection about the appropriate techniques would be useful. There is a large range of computational techniques we might choose from, which calls for reflection about that choice. We need not only to apply techniques, but also to justify the choice of technique, in order to develop a sound and transparent methodology.

Third, and finally, and quite within the spirit of considering interdisciplinary work, I would like to remind the linguists and dialectologists here that we are dealing with space at nearly every turn. This suggests that we should examine and consider how other disciplines have analyzed spatial relations, e.g. human geography. Human behaviour produces spatial consequences and constructs. We ought to examine the techniques applied there with an eye on opportunities for learning from each other.

***David Robey** asked what scholars using computational techniques in the various Humanities disciplines may learn from each other. His remarks follow.*

My focus here is on the institutional and strategic context of humanities computing in the UK, and particularly on the recent changes in that context. The focus on the UK will also be of interest to researchers in other countries, I hope, because we have recently moved from a position in the digital humanities that was exceptional, in international terms, to one that presents much the same kind of problems as that of other countries.

Until 2008 the UK had the strongest system of support services for ICT use in arts and humanities (A&H) research in the world. This was made up of, first, the Arts and Humanities Data Service (AHDS), joint-funded by the Arts and Humanities Research Council (AHRC) and the Joint Information Systems Committee (JISC), to serve as a national repository for data outputs in the A&H, and as a source of advice on data creation, curation and preservation. This was divided into a number of Centres, of which the most relevant for linguistics was AHDS Literature and Linguistics, associated with the Oxford Text Archive at Oxford University Computing Services. Funding for the AHDS was terminated at the end of March 2008, largely as a result of financial exigencies at the AHRC. The AHRC also argued that the use of the AHDS's resources did not justify the considerable cost of maintaining it, and that much of the expertise that it provided was now widely available in UK universities.

A second major source of support was the AHRC Resource Enhancement Scheme, the last grants from which were made in 2006. This funded mostly digital resource outputs, overall a total of some 175, to a total value of almost £40m. Some examples in the area of linguistics and dialectology are: Yaron Matras (University of Manchester), *Morphosyntactic typology of Romani*; Karen Corrigan (University of Hertfordshire), *The Newcastle Electronic Corpus of Tyneside English*; Justin Watkins (University of London), *Wa dictionary and internet database for minority* *languages of Burma*; and databases of the Scots, Irish and Welsh languages. To see the AHRC's review of the Resource Enhancement Scheme, visit

http://www.ahrc.ac.uk/FundedResearch/Pages/ResourceEnhancementSchemeReview.aspx

The third and most recent source of support was the AHRC ICT in Arts and Humanities Research Programme, funded to a total of £3.8m from October 2003 to December 2008: for details see http://www.ahrc.ac.uk/ict. Its aims were to build national capacity in the use of ICT for arts and humanities research and to advise the AHRC on matters of ICT strategy. Its largest single project was the AHRC ICT Methods Network, a forum for the exchange and dissemination of advanced expertise on the use of ICT for A&H Research, which ran for three years to the end of March 2008. It also funded a set of ICT Strategy Projects, a mix of survey and tools development work, of which two are particularly relevant to linguistics: Alan Marsden (Lancaster) and John Coleman (Oxford), *Strategies, Requirements and Tools for the Analysis of Time-based Media*; and Jeremy Smith (Glasgow), *Lexical Searches for the Arts and Humanities*, which developed a new user interface for the Glasgow Historical Thesaurus of English. Its other major activity is the Arts and Humanities e-Science Initiative, funded jointly with JISC and the Engineering and Physical Sciences Research Council, which has extended to the A&H the e-Science agenda for the development of advanced technologies for collaboration and resource-sharing across the Internet. While it may seem oxymoronic, at least in English, to speak of e-Science in the A&H, the e-Science agenda is important for the A&H above all in its approach to the problem of data dispersal: a major problem that confronts the humanities researcher is not the lack of relevant electronic data so much as the fact that so much of it is dispersed in self-contained databases and archives that do not communicate with each other; yet it is a truism that the more data can be connected with other data, the more it tends to be useful. For more details on the joint Initiative, see http://www.ahrc.ac.uk/e-science.

What then are the problems that confront us now that we have come to the end of this period of exceptionally strong support for ICT in A&H research? The first is that of finding secure repositories for the publicly-funded data outputs that would previously have found homes in the AHDS. While the former AHDS Centres still remain in existence, supported by their host universities, and still maintain access to the collections that have been deposited with them in the past, it is far from clear how much longer they will continue to do so, nor are they in a position to receive new deposits without some form of special funding. In the longer term a national system of institutional repositories (IRs) in universities may provide a solution to the data preservation problem. There is a strong public agenda to develop such a system, largely driven by JISC, but for the moment the IRs that exist serve mainly to preserve open-access e-prints of journal articles, and very few are capable of handling other types of data. A publicly-funded study is also currently under way on the feasibility of establishing a UK Research Data Service (see http://www.ukrds.ac.uk/), which may eventually provide a solution to the data storage problem, but whatever the outcome of this study is, it is unlikely that the solution will be effective in the short-to-medium term. In the meantime a good proportion of the Resource Enhancement projects funded by the AHRC have serious concerns about the long-term sustainability of their outputs: see the recent review at

http://www.ahrcict.rdg.ac.uk/activities/review/sustainability.htm.

But the most urgent issue is to maintain the knowledge and expertise that was built up over the years by the AHDS and, more recently, by the AHRC ICT Methods Network; and together with this to keep in existence and develop the communities of practitioners that built up around the two organizations. The Methods Network in particular, through a series of some 50 expert seminars, workshops and similar meetings, was remarkably successful in bringing together advanced practitioners of digital methods in the A&H, many of whom were previously unaware of each other's existence, and most of whom had previously had little direct contact with one another. It is vital, if we are to maintain the current quality and volume of UK activities in the digital arts and humanities, to keep these communities in existence—a point particularly relevant to the topic of this session.

Support continues to be provided in a limited number of areas. The AHRC still funds the former Archaeology Centre of the AHDS, on the grounds that the use of digital resources is more firmly embedded in Archaeology than in other A&H disciplines. JISC continues to provide direct support for e-Science activities, now broadened to include all advanced ICT-based methodologies in the A&H, through its A&H e-Science Support Centre; for details see http://www.ahessc.ac.uk/. It also needs to be emphasized that provision of generic ICT infrastructures for research remains strong, mainly through JISC, but it is not part of JISC's mission to provide for individual discipline domains such as the A&H. Support for research projects using ICT in the A&H also continues to be strong: a high proportion of the standard research grants made by the AHRC under its responsive-mode schemes continues to include the production of digital outputs of one kind or another. The important issue is not the provision of the generic e-Infrastructure, or the support of ICT-based research projects, but ICT support services, resources and infrastructure specifically for A&H research: the A&H e-research infrastructure.

In current circumstances, and pending a possible solution of the data repository issue, the best immediate prospect is to develop support in virtual mode. With the help of transitional funding from JISC, considerable effort is currently going into the development of a set of on-line resources for the use of ICT for Arts and Humanities research, based at the Centre for e-Research at King's College London. These will maintain a knowledge base of training and methodological materials, and at the same time support virtual communities of practitioners, including a network of expert centres in data creation, preservation and use. The knowledge base will support the virtual communities, who in turn, it is hoped, will feed material into the knowledge base: see http://www.arts-humanities.net. By these means we hope to maintain and build on part of the legacy of what, with hindsight, was a quite exceptional period of public funding for digital A&H support services. But it is hard to see how, without further public investment, we can maintain the current volume and quality of activity in the longer term.

Discussion

John Nerbonne: Thanks to our panellists for their opening remarks. I'd now like to open the general discussion by asking if the panellists have points of discussion or questions for each other. For example, David addressed one of the things that Paul had said earlier, namely that the value of the long-term preservation of digital resources for the humanities is still unsure.

David Robey: Paul, you didn't quite explain the value of the resources in historical linguistics, did you? Of course, you did explain what you are doing and why.

Paul Heggarty: The value for others in the discipline is certainly clear in the case I described, but how it’s perceived depends in general on the orientation of the other people in the discipline. Most disciplines will have a divide between the computationally inclined and the non-computationally inclined.

John Nerbonne: Maybe a question then with respect to the point about the perceived value of the computational work: you opened your presentation, David, with the pessimistic view, or at least pessimistic report, that the very good level of funding in the UK will not be continued, while we have heard from both Paul and Roeland van Hout that they see that the computational work in the Humanities, the interdisciplinary work involving Computing Science on the one hand and Humanities on the other, and sometimes disciplines outside the Humanities as well, has been really quite successful. Has there been a final verdict in the UK then, that the time when you needed to fund very special computational efforts in the Humanities has passed, and that it now should become a more normal way of working in the Arts and Humanities, one that requires no special attention and no special funding? I didn't know about the report, and your presentation surprised me.

Roeland van Hout: In the Netherlands we see a version of the funding problem that is particularly frustrating. Everyone is looking to one another, asking who is responsible. Everyone knows that money needs to be invested, but whose budget should it come out of? For example, the money in the Netherlands from our National Science Foundation needs to be invested in projects. So long-term grants are not given at all, and projects requiring long-term budgets should be taken over by some other institution. Alternatively, the results of the projects become the responsibility of some institution, and that is also a problem in the Netherlands. When working with short-term (3-4 yr.) projects, it means that at a specific moment there is a lot of money, and suddenly later the money is gone, and you have to look for other funds, making the sustainability of the whole enterprise a problem. Not only in the Humanities, by the way, but also in other branches of science, too, where there is actually more money than in Linguistics and the Humanities (but that's another problem).

John Nerbonne: Linguistic searches and frequency counts look like quick and easy results that need no special funding. Maybe we’ve been too successful in showing how much benefit accrues to even modest computational efforts.

David Robey: Research councils are fond of quick results.

Paul Heggarty: I wonder also whether there’s a cross-disciplinary divide in how these things are perceived. So if you look at publications on Linguistics in the big natural sciences journals -- *Science*, *Nature* or the *Proceedings of the National Academy of Sciences* (PNAS) -- they are quite reasonably seen as important because they have been reviewed and accepted in these very selective journals, even though colleagues in the Humanities have often been very critical about their content. There’s also the attitude reported by some colleagues who work with natural scientists, that the latter, the natural scientists, feel a sort of frustration with the computer literacy of people in the Humanities. The responses to the PNAS article I mentioned involved some terrible mud-slinging across the disciplinary gorge, where Forster and Toth were really slammed for not doing linguistics properly. It was said that they really did it very badly, while Forster's response in an interview was along the lines of: "Well they [i.e. linguists] just don't understand what we are doing". There are a lot of traditional people in the Humanities who don't see the value of computational work. Unfortunately, they often don't know what's going on; for example, they understand the statistics so poorly that they're very suspicious about its value. I just wonder if this distrust came through in the debate about funding.

David Robey: This is quite interesting. Much of the funding from my research council has computational aspects, and in the latest, most recent rounds of research grant applications about half of the projects involved some kind of digital output. In many cases the applications were not required or even encouraged to plan digital outputs, but the community submits applications with digital aspects anyway. There are other indications as well that information technology is actually a fairly fundamental part of the way in which about half the community works. There is quite a lot going on, I think. I think the real problem is for it to be done more effectively. That's the thing that worries me, that it's not always done nearly as effectively as it should be.

Roeland van Hout: John, don't you want to take comments from the audience, as well?

John Nerbonne: It sounds like a very good time to get our audience involved. If you wouldn't mind using the microphone, that would help. I need to record you because I want to make this available later.

Eric Wheeler, Toronto: My comment concerns our talking about sustainability, and how you have to go after follow-up grants after your first grant is up. I'm worried what happens about 100 years from now, when I may be retired and not able to go after grants. I still want the work I am doing to survive. If we look back 100 years or 200 years ago, the tradition was that you wrote a book, and it was put on the library shelf. Now that library will be there forever, but when you create a database, you put in on the shelf, and you hope that your software isn't obsolete before the first colleague asks you for a copy of it.

John Nerbonne: And it's not only that your software becomes obsolete, the entire environment is changing all the time.

Eric Wheeler: Well, it's the environment changing, and the demands, and the directions, and there are lots of things, but we don't have this institutional sense of how to preserve, use, support, maintain, and distribute what we are producing digitally. This is related to a second problem that I'd like to mention, namely that we are still very much pioneers in this, even though we may be jumping on a band-wagon that started long ago. I have worked across disciplines I guess for my whole career, and I have always been put down for not being in the right discipline. I am always at the margin of something, and when the margin is information technology, and you see everybody using it, you say "This isn't right somehow, how can my work be viewed as marginal?" I mean they should come begging me to do this work, and they don't. And that is another fundamental problem we have, I think. How do we make it a priority to solve these problems? They are huge, long term problems that need big money and big thinking and more than just one individual coming up with one individual solution. But the rest of the Humanities depends on us because everything you discover sooner or later you are going to want it put in digital form and so that we keep it.

John Nerbonne: Paul, want to pick up on that?

Paul Heggarty: As an anecdote on preservation problems, the 2003 article in *Nature* by Gray and Atkinson was based on a database by Dyen, Kruskal and Black of 90-odd Indo-European languages that had originally been entered on punch cards, I think. Dyen and his co-authors later re-entered it, making it available at a website at a university in Australia where one of them [Paul Black] worked. But as of two years ago, the site was gone. I've done searches at the university, and you just can't find it anymore. I happen to have a copy, but I don't know where it's available anywhere on the internet. That was a huge dataset of 93 languages, 200 words per language, used for a big article in *Nature*, that caused a huge amount of fuss. Now nobody knows where it is. [Update: after the debate a member of the audience suggested I contact Paul Black directly about this. I did so and eventually another site was found that is now hosting the database.]

John Nerbonne: Bill Kretzschmar, who makes lot of data available to people, wishes to contribute.

Bill Kretzschmar, Georgia: I try. Well, one piece of good news on this front. My university librarian tells me that it is a big topic in conversation among the librarians in America. I've actually made an agreement with my university library to store my archive, which is, well, large. It is in the range of 20 terabytes of material that they are going to store and maintain as part of their digital media, multimedia archive, and what I contribute is only a tiny piece of it. I want to store TV, movies and all sort of things that are very space-intensive, and so I think that this is maybe a movement we need to take interest in, which means that we need to be talking to our university librarians. To me it always seemed odd that the university was willing to take indirect costs for research grants but was unwilling to archive the results of those grants. I think now is the time for us to be active on that front.

David Robey: I think that's absolutely right. Libraries are the obvious place to look, even if librarians sometimes need quite a lot of persuasion. That's also a key solution, since the money isn't coming from anywhere else.

John Nerbonne: Are there further discussion points?

Fiona Douglas, Leeds: A question really, I suppose for David, involving the same point I made when there was the recent consultation process on whether we should have a national repository or not, and what it should look like. It seems to me that having a Humanities-wide approach might mean that we avoid the current problem where we have lots and lots of data that might be really useful to other people who might use it in entirely different ways. Maybe the way to market the more general approach is to note that the data might be useful to other people. So, for example, you know, I have worked with newspaper texts, but all the newspaper resources are designed for journalists. They are not designed for linguists and yet lots and lots of researchers use newspapers as, you know, examples of standard English around the world. It seems to me crazy that we spend all this money setting up resources for only one very specific niche market when with very little effort we could extend that and make it much more available and more generally useful to other researchers with very different interests. I mean many of you might be interested in oral histories not because you are oral historians but because you are interested in the language that is used. So, especially if the research money is increasingly limited, so that we have to kind of make the resources we have go further, one way of doing it would be to stimulate resource builders to make their resources more generally accessible and more generally useful.

David Robey: Well, I think Fiona has put her finger on a very important point. In this country we are probably not going to have enough funding for a specialized Arts and Humanities data service again. I would guess that it is not going to happen. So what we have to do is to find some kind of infrastructure that can deliver that data at lower cost, working as a default with the university libraries and such. I think that's the agenda for the future.

Roeland van Hout: There's also another point here where I think we researchers also can do things. 10 years ago when one was collecting data, the data were personal data, and there was no intention to ultimately make the data more generally available. I think that nowadays we should not accept, for instance, articles written on the basis of a database if the database is not available. Further, making the database available means that it has an open structure, so not an idiosyncratic, unstructured form. Here I think that researchers themselves can make the information more effectively available so that databases have much better chance to survive. And that was the point of your argument.

David Robey: I think people need to think about publishing databases in the same way they think about publishing books.

Patrick McConvell, Australian National University: Can’t the research councils make it an obligation when they fund a grant hosted by a university, that the university has to look after the results and provide access to the data. Wouldn’t that be possible?

David Robey: In real life I think probably not. I mean at the moment my research council requires three years’ maintenance after the end of the grant period, and there is a question about whether that can be extended. But of course we don't ever have much control because the money has been spent, and then influence is over.

Paul Heggarty: On that cross-disciplinary point I remember something that Russell Gray said. He was trying to get hold of language data to compare genetic and linguistic groups. He found it incredibly difficult to obtain the linguistic data as opposed to the data from genetics. The human genome project has resulted in databases with world-wide standardizations for genetic information, and he just couldn't believe how bad linguistics was when it came to reliable standards.

Eric Wheeler: It is as much our fault because we do it in craftsmen-like, small-scale ways, and we need to do it on a larger scale. I mean the fact that we have these databases and they have names is wonderful, I mean that's a vast improvement over what was there 10 years ago. But it’s not quite the same as the Human Genome Project, and we wonder why we don't get funding: it's because we look like poor cousins when we come in with our projects, the poor cousin projects that get poor-cousin funding.

Bill Kretzschmar: Part of the reason for that is that we are a diverse field in language studies, that we are not all doing the same thing or close to the same thing, and that the datasets don’t look the same. We have radically different datasets and radically different uses we make of them, and I just don't believe that we can have one style of encoding that works for all of us. SGML tried to do this, and TEI has made a stab at language, too, and failed, as far as I'm concerned, but I think that’s just because it's an impossible task. We will always have divisions. We have to accept that.

David Robey: But you can add degrees and degrees of reusability to that. You can improve reusability, but you can't have total reusability.

Bill Kretzschmar: We can't throw things away just because we have different encodings. We have to preserve the metadata so that people can use our encodings.

John Nerbonne: Right, I also see an enormous reuse of the data that is available. I don't know how many articles have been written about the Wall Street Journal (Penn) Treebank and the different sorts of syntactic analysis that have been applied to it. One group at Penn took the time to syntactically annotate it, and surely there have been 80 or 90 articles on it, it's really enormous. There is hunger for data. And for comparable analyses, using the same dataset. Reusing data is not merely efficient, it often allows you to say a lot more.

David Robey: Actually I think it's relatively easy for you guys to demonstrate evidence of value compared to history and literature studies where that's so much more difficult.

John Nerbonne: It's a good point that you brought that up because actually we look at the computational side of things most of the time. Perhaps that's what gets the imagination going here. What about the cross-disciplinary aspects, one of the perspectives we wished to address in this discussion? Paul most explicitly addressed interdisciplinary issues. Historians, archaeologists, geneticists are all interested in some of the same questions, and therefore interested in collaboration. What about the other Humanities disciplines? What about History and Religion? Anthropology? Are we witnessing increased collaboration through computational studies in these other areas of the Humanities?

David Robey: There's a lot of use going on of geographical information systems, spread across a surprising number of Humanities disciplines. Among the fields you see using GIS are History, Archaeology (obviously), and Literature studies, sometimes.

Eric Wheeler: I think again one of the concerns is that GIS systems were designed for geography and are being adapted piecemeal to other things. If we really got the experts in geography to come and look at what our problems were, I think they would design the systems differently. Our personal experience in that was thinking that when we developed our dialect software, the mapping was going to be hard, but it turned out to be very easy. What was hard was the interface to access the data because we asked different questions than you would typically ask in geography or anywhere else. And I think that is going to happen all across the Humanities. It is understanding what my data is all about and what the interesting questions to ask are about a dialect or a language or whatever, that will drive that the developments. But there is still a need for people who have expertise to develop systems. They can help a great deal. A meeting ground for these groups would be useful.

David Robey: I think there is a larger sustainability problem with interface software. Online data is much easier to preserve. – Just keeping online searchable interfaces sustainable is a real problem, not to mention creating them.

Eric Wheeler: And even having a standard for what they should look like. I mean much as we curse Windows, we now have a point where we can talk to files and systems in a fairly standard way. Wouldn't it be nice if you could ask your dialect questions in some kind of standard ways?

John Nerbonne: Generic tools?

Eric Wheeler: Generic tools.

Bill Kretzschmar: I don't believe that they’re feasible, at least not beyond very simple programs.

Eric Wheeler: I agree that it is difficult.

Bill Kretzschmar: The point is, we have to steal everything possible from other people who already made them. But the creation of generic Humanistic tools is something that people always talk about in the digital Humanities, and that I'm never been convinced of buying into.

Roeland van Hout: No, but, it's my experience that there is a lot of information around nowadays about all kinds of cultural and historical things, which can be related to the development of dialects for instance. It is admittedly sometimes problematic to get this data. For example, all the institutional borders, political borders, etc. It's not very simple but it has become easier to combine the linguistic data with many other types of data, and that's maybe a message we can have for the Humanities. We can use all these different resources now!

David Robey: That's why we got excited by e-science. Because that's exactly what e-sciences are about: about integrating data. – But they, too, have a long way to go.

Roeland van Hout: Maybe not that long, but OK.

Bill Kretzschmar: One good thing about this is that in a long period of time seeking research funds, I've never yet been successful in getting money to build a tool. I think that funding organizations are resistant to people who want to build tools, and they would rather have data collections or innovative methods. Don't set out to build a generic tool, but look instead for a breakthrough in analysis.

Patrick McConvell: We just had a one-year experience in Australia with money that was given out for research in the Humanities, precisely to build tools. But it lasted for one year maybe, and then, ... then, the money ran out.

David Robey: I mean if there’s a problem with the sustainability of data resources, then probably the sustainability of tools is of a different order, isn't it?

John Nerbonne: It's true but there are examples, well known in the community of variationist linguists. The VARBRULE package, which was built when one could only use 4 K on a PDP 7 somewhere. It has been around for 30 years, which is amazing. People are still using it. David Sankoff really deserves credit for this.

Eli Johanne Ellingsve, Oslo: I am one of the editors of the Norwegian dictionary in Norway, planned to be finished around 2014. We have started discussing connecting with historical databases, Archaeology, Geology, sites of interest for nature research, including petro-geology (oil), and recently the ordnance service. (Is that what you call it in English? -Yes.) -The mapping authorities hold a key position. Most importantly, they organize a great deal. In Norway they are well organized, and we plan to connect to their structure and their data system. We have started discussions with them, and there is interest. They are state funded, i.e., officially funded. So they have money, and they would like us to connect to build up the linguistic databases, to provide more information, including linguistic and historical linguistic information as well. So the ordnance service in different countries may be as interesting as it is in Norway. You may try that.

John Nerbonne: Thank you. I'm going to make a round of the panellists now, because we have a plenary talk at 16.45 and that'll give you just about enough time to get there. But maybe we'll go in reverse order this time. David, do you have any final remarks?

David Robey: I'll pass, as I've talked quite a lot, thanks.

John: Ok, then Roeland.

Roeland van Hout: My final remark concerns the state of databases, but again especially the interaction between institutions and the producers of the databases. So I would like to repeat: if you are creating databases, try to make these databases available and take responsibility for your own data, because that may be helpful to convince the larger parties to take part in our enterprise.

Franz Manni, *Musée de l'Homme*, Paris: May it also be useful to print them sometimes? If they are printable.

John Nerbonne: OK, and let's give Paul the final word, then.

Paul Heggarty: Sometimes I think, maybe I should put my data on Wikipedia or something. It's the best place I know of -- because no one else is doing it -- to get Swadesh lists for hundreds of languages. It's not on Wikipedia itself, it's one of its sub-branches [the Wiktionary Appendix system], but the lists are pretty much standardized. I use Google for my language mapping as well. These are things that are becoming *de facto* standards. And it seems like linguists may be forced to work via Wikipedia just because academic organizations are not doing it anywhere. It's a bit of a shame that we have to wait for “Google Lingo” or some such to come out... But at least the Wikipedia sites use some sort of standardization. Maybe solutions are arising through the web and through organizations like Google or Wikipedia.

John Nerbonne: OK so we close the session with a word on the web. We thank in particular our panellists!