AHRC ICT METHODS NETWORK

AHRC ICT Methods Network Workshop

USING LARGE-SCALE XML CORPORA IN LANGUAGE AND LITERATURE

OXFORD UNIVERSITY COMPUTING SERVICES, 26 NOVEMBER 2007

Report by Lou Burnard

Introduction

Since its first release in 1994/5, the British National Corpus (BNC) has become a key resource for researchers, learners, and teachers in English language teaching, linguistics, Natural Language Processing, lexicography, cultural studies, and many related fields. It remains amongst the best known and most frequently accessed resources of its type worldwide. In March 2007, a new edition of the corpus was released in XML format. The decision to convert the corpus into XML was based on a number of factors: XML is increasingly the standard for online text creation and publication; tools for processing XML resources are ubiquitous; other linguistic resources comparable to the BNC are increasingly created using XML. Converting the BNC into XML thus improves its usability by making it possible for users to access it with their own tools, drawn from a wide range of new sources, and to integrate it with other resources.

Despite its wide take-up on the internet, XML remains less well understood by researchers and resource users from a non-technical background, who may therefore find it difficult to identify or make use of existing information about how to benefit from the opportunities available when using XML.

This one-day workshop therefore aimed to introduce the technologies needed to unlock the potential uses of large-scale XML-encoded language corpora, with a particular focus on the BNC XML Edition. The workshop was aimed at two distinct groups of researchers. The first group contained language or literature specialists who are aware of the potential for corpus-based methods in language pedagogy or literary research and want to apply them either with their own corpus material or with the BNC in its new format. The second group was made up of technical specialists who are aware of the demand for corpus resources and wanted to gain practical experience of using XML for corpus creation, development, and usage. Through the workshop we were hoping to stimulate dialogue between the two groups, and promote a shared understanding of common goals.

The workshop was advertised on the BNC home page and a number of mailing lists (e.g. Corpora, Linguist List, TEI-L), and via the HEA Subject centres as well as through various specialist groups such as the BAAL corpus SIG. The number of applicants was far higher than the number of places available; in fact the waiting list was as long as the list of participants. The workshop attracted applicants from across the UK, Europe and as far away as India, USA, and South-Africa, demonstrating the international importance of the British National Corpus and the worldwide interest in the development and use of large-scale corpus resources and XML.

The workshop was organized as a series of sessions, each including a presentation and a practical component. The material used in the sessions (presentations, hand-outs and exercises) was made available online after the event, something that was deemed particularly important in light of the number of applicants that could not be allocated a place.

Session 1

The day started with an introduction to the BNC, focusing in particular on the ideas underlying its conscious design principles, and the social, theoretical, and technological context in which the corpus was created. The presentation described the principles underlying selection of material for inclusion in the corpus and how its creation was organized and managed as a project. Because the BNC consists of a large number of individual text samples, an understanding of the selection principles is necessary to exploit the full potential of the resource, as is a grasp of the way information about individual sample texts is presented.

The BNC texts were selected on the basis of carefully pre-defined criteria such as publication type, year of publication, and text domain. Further information about the texts was also recorded when available, even if it was not part of the selection criteria. The corpus consists of a corpus header file, containing information and metadata relevant to the whole corpus, and a series of corpus texts. Each corpus text consists of two parts: a header containing information and metadata about the individual text, and the text itself. The texts contain information about textual features (headings, page break, etc) and features specific to spoken material (speaker turns, change in voice quality, non-verbal events etc). At the word level, every word is annotated with word-class and lemma. The XML format used to record all of this information was presented informally. The slides used for the presentation are available via the workshop page on the BNC website: http://www.natcorp.ox.ac.uk/workshop/26nov.xml.

In the practical session following this presentation the participants had a chance to explore a BNC XML text 'in the raw', studying the structure and mark-up with different kinds of tools. They learned how the texts could be displayed in a text editor which could not process the markup in anyway, or in a web browser which could display it under control of a stylesheet. Using an XML editor (Oxygen), the participants then explored how the texts could be displayed in other ways, for example without the markup, with the information in the markup converted to a more reader-friendly form, or in a format suitable for use by some non-XML aware software. The stylesheets and other resources used here are also available from the workshop website.

Session 2

The second session of the day began with an invited lecture from Professor Guy Aston (University of Bologna) discussing the use of the BNC XML Edition for language teaching and learning. He argued that the use of online corpora like the BNC can be empowering for both learner and teacher, if approached in the right way, and discussed several specific examples of the kinds of exploratory questions that he had found useful in his experience as a teacher of translators and interpreters. Tools such as XAIRA stimulate exploration, and often leads to serendipitous insights about the language. His talk included screen-shots from the XAIRA program to illustrate how particular queries were formulated, and is available via the workshop webpage: http://www.natcorp.ox.ac.uk/workshop/26nov.xml/gaBNC_ox07.ppt.

This presentation was followed by a practical component where the participants were given a set of tasks to solve using XAIRA and BNC XML in the manner described in the lecture. Each task introduced a different feature of the XAIRA program and illustrated how an awareness of the structure and coding of the corpus affects the way you formulate a query. The tasks were organized in an informal way, representing the way in which resolution of one question leads to formulation of the next: for example, exploration of the different spellings for 'hard-boiled', 'hard boiled', and 'hardboiled' (all attested in the corpus) leads to such questions as 'what kinds of thing are commonly described by this phrase and in what kinds of text'?

Session 3

As noted above, one of the aims of the workshop was to stimulate dialogue between different kinds of researcher. This was achieved both informally during breaks and over lunch, which was provided, and also

in a special session during which participants were invited to present the kind of material they were working on, and to ask for and offer feedback, suggestions and comments. This discussion led to a wide-ranging and open-ended discussion of key issues relating to corpus design, the special problems of spoken corpus creation, permissions issues, and the practical limits of corpus annotation. Participants included a mix of technical and language experts (with the latter predominating), which led to some mutual incomprehension, but did not impede very good humoured debate.

The discussion session was followed by another scripted practical session in which participants explored the indexing function in XAIRA, which makes it possible to use this tool with any XML-marked-up corpus, including participants' own materials. For the exercise, three versions of a literary corpus were made available: a plain text version, a version in XML format with minimal mark-up and a version where the corpus had been annotated with word-class information (in XML). The participants used the Indexing Wizard to index each of these three versions and learned how the different XML tagging in each version affected the usability of the resulting indexed corpus. Detailed instructions for this exercise and the sample corpora used are all available from the workshop website.

Session 4

In the final session of the day, we tried to generalize the tools and techniques presented earlier, by discussing how they might be applied to in the creation and exploration of different kinds of corpus. Guy Aston described his work on his ‘Any Questions’ corpus, a fascinating collection of speech transcribed from the popular radio show. He showed how some research questions about male and female speech patterns, and about differing contexts and political attitudes might be explored by means of this corpus, and also discussed some of the methodological issues it raised, notably those relating to the collection of such material, questions of copyright clearance, transcription practices, formatting and annotation. These topics connected well with issues raised earlier and indeed throughout the day.

In the last practical session of the day, participants were given the opportunity to focus in more depth on one of the topics covered in the workshop. Some of them chose to explore the XAIRA tool further, working on their own material; others to explore some of the alternative corpus interfaces available for the BNC; others to discuss their specific research questions with one of the workshop teachers. As a final round up session, Martin Wynne chaired a review of the day and of the key issues it raised for the participants.

Conclusion

The workshop received a considerable amount of attention and interest, which shows that there is a need for training of this kind. The fact that it was advertised mainly via linguistic channels may account for the type of participant it attracted. Considering that the recruiting ground was narrow and relatively specialized, the event still attracted more people than could be accommodated. A similar event has since been re-run to cater specifically for applicants from the local institution, with a limited number of places available to external applicants. The fact that that event also filled up very quickly is a further illustration that the demand for this kind of training is not yet exhausted.

The evaluations on the day were very positive, and participants also contacted the course team after the event to express their appreciation. Among the suggestions that came up, and which could form the basis for future events, was that more time for practical, hands-on work would be appreciated. To accommodate this hands-on work the workshop could be re-run as a two-day event. The fact that participants voluntarily spent extra time on the practical tasks during breaks and lunchtime and at the end of the day offers further support for this idea. Several participants (at this event and at the local repeat session) also expressed a wish to learn more about corpus creation and annotation and to get training in the practical aspects of creating your own corpus.

This enthusiasm testifies to our belief that the exploration of large scale textual resources within the corpus linguistics paradigm constitutes a very important research method in itself, of central importance in the domain of language studies and language learning, but also more widely in literary or social studies. The techniques appropriate to the automatic analysis of large scale linguistic resources are not intuitively obvious to researchers in the humanities, and even amongst those engaged in the creation of digital textual resources, there is often a lack of awareness of what can be done using them beyond simply reformatting them for display on screen or on paper. Such techniques have clearly proved their worth in the teaching of language, and have recognized potential within the Humanities more widely. Corpus linguistics as currently understood is, after all, one of the few humanities disciplines that can only be done by means of information technology.

Workshop website

Full information about the workshop, including the timetable and all materials presented, both at the Methods Network funded session, and the local event mentioned above, are available from the workshop website at http://www.natcorp.ox.ac.uk/workshop and the Methods Network activity page: http://www.methodsnetwork.ac.uk/act30.html.