Workshop Report: Corpus Approaches to the Language of Literature

Authors: Ylva Berglund and Martin Wynne

INTRODUCTION

The *Corpus Approaches to the Language of Literature* workshop took place at Oxford University Computing Services on 17-18th May 2006. The event was the first in the series of advanced workshops funded by the AHRC ICT Methods Networks, and it gathered more than 20 participants from different geographical areas, research backgrounds, and subject fields to attend a series of presentations and practical sessions over a two-day period .

The event aimed to disseminate advanced methods in linguistic analysis using linguistic corpora to researchers in literary studies.

The workshop built on networks and discussions at a workshop on Corpus Approaches to Literature held at the Corpus Linguistics 2005 conference in Birmingham and recent Poetics and Linguistics Association (PALA) conferences. These discussions resulted in a clear feeling that while there was a recognition of the potential usefulness of corpora, there are practical barriers to progress. It was decided that it would be useful to run an event which would (a) disseminate examples of exemplary work in the field, and (b) introduce in a practical way literary scholars to the techniques and methods of corpus linguistics. The workshop was supported by the Poetics and Linguistics Association, and access to their email list and website proved to be a very useful way to publicise the event to a key constituency. The workshop was the founding event for a PALA special interest group on corpus stylistics, and a short report was published in the *Parlance*, the PALA newsletter.

The workshop looked at practical ways to exploit the potential for more widespread use of corpora to study literature. Work in stylistics relies on the evidence of the language of literature. Corpus linguistics is also an empirical approach to linguistic description, relying on the evidence of language usage as collected and analysed in corpora. As linguists and stylisticians become more aware of the possibilities offered by corpus resources and techniques, then a useful exchange of ideas and methods can be facilitated.

The workshop was an opportunity to disseminate and discuss examples of successful research which has shed new light on literary texts through the techniques of corpus linguistics. Furthermore, it pointed to ways forward in demonstrating the resources and techniques necessary for such work in the future. Participants were armed with arguments, language resources, tools and methods to take back to their departments to train colleagues and to use in their research and teaching.

Discussion addressed the following topics:

(i) the study of literary effects (or 'deviations') in texts by using the evidence of language norms in a reference corpus, including the use of collocations, colligations and semantic prosody;

(ii) creativity in language, as identified or analysed with reference to corpus evidence;

(iii) corpus annotation and analysis as a means of conducting a thorough and exhaustive analysis of linguistic features in literary texts;

(iv) theoretical and practical problems with the use of corpora in literary study;

(v) resources and techniques for the study of literature using corpora.

The techniques which were explored are also of use in teaching literature and linguistics. It was therefore possible to work with the HEA centre for English to publicise the event. HEA English expressed strong support for the venture, and a representative was present at the event.

Participants were asked to provide a short statement indicating the area of work relevant to the workshop in which they were involved, and explaining what they hoped to get out of the workshop. This information is provided as an appendix below.

REPORT ON THE PROCEEDINGS

The event started with a short introduction to the area from Martin Wynne (AHDS Literature, Languages and Linguistics), where it was suggested that this is a relatively new field, a field where corpus linguistic methods meet and mingle with literary analysis and stylistics. The introduction was followed by a practical session where the participants had a chance to explore some key tools and techniques as well as the resources made available.

Jonathan Culpepper (Lancaster University) then described how techniques developed in corpus linguistics can be used to produce a new kind of dictionary based on usage. With illustrations from a number of case studies, he showed how he used familiar notions in corpus linguistics, such as collocation, cluster (multiword unit), keyword and grammatical and semantic annotation to examine the language of Shakespeare.

Michaela Mahlberg (University of Liverpool) further developed the notion of ‘corpus stylistics’ as a meeting between corpus linguistics and literary stylistics, stressing that it is not simply the application of corpus linguistic methodology to the study of style. With illustrations from her studies of the language of Dickens’, she showed that if we use innovative categories to describe linguistic norms, deviations from these norms will shed new light on the way in which we analyse style in literary texts.

The third presentation was by Bill Louw (University of Zimbabwe). In his talk on ‘Collocations, corpora and criticism’ he provided novel illustrations and inspirational examples of how to look at collocations when examining literary works. He suggested that ‘collocation has begun to offer proof of its ability to produce tangible results which exceed the results of close reading and far outstrip approaches fettered by grammar and syntax alone’.

The practical, hands-on sessions were an important part of the workshop. Each presentation was followed by a practical session where the workshop participants were given an opportunity to explore the methods used by each presenter. The sessions were structured to allow a natural progression, from simpler to more complex methods and they were followed by general discussions where problems and success stories were shared. To allow not only the participants to benefit from these, the workshop material (abstracts of presentation, exercises and practical guides) will be made available on the workshop webpage (http://www.ahds.ac.uk/litlangling/events/approaches/home.htm) and via the Methods Network webpages.

OUTCOMES

The workshop was viewed as a success by all organisers and participants, and excellent feedback was received, both through the feedback forms completed by participants, and the many votes of thanks and encouragement received verbally and by email, which still continue to arrive.

This event led directly to a one-day pre-conference workshop at the PALA conference in Joensuu in Finland, at which more PALA members and other international scholars were able to participate.

The exercises and presentations are currently online and freely available. Work is ongoing to improve the accessibility and sustainability of the online resource.

The workshop was successful in linking and 'joining up' various services supporting academic work in the UK, via the involvement of AHDS, Methods Network and HEA. The workshop was also successful in raising the profile of these services and in building links between organisations and communities in linguistics, literature, stylistics and humanities computing. The workshop itself only lasted two days, but the aspiration is that it will be the beginning of future fruitful collaboration and exchange of ideas. Among the possible initiatives that were suggested was a special group for new post-graduates in the field, and the development of some materials to exploit the pedagogical potential of some existing resources. Participants who were interested continuing discussion of the issues raised in the course were invited to join the 'corpus-style' email list (http://www.jiscmail.ac.uk/lists/corpus-style.html).

APPENDIX: STATEMENTS OF INTEREST BY THE PARTICIPANTS

Name: Jarle Ebeling
Oriental Institute, Oxford

[No information given, but he is known to be building a corpus of ancient Sumerian]

Name: Koldo J. GARAI
Email address: KoldoGarai@terpalum.umd.edu / Koldo.Garai@laposte.net
Institutional affiliation: IKER - Basque Text and Language Study Center (CNRS) Bayonne

I heard about the event at the PALA web site.
After several years of teaching in several institutions (University of Deusto, University of the Basque Country, University of Bordeaux III Michel Montaigne) I am currently finishing my Ph.D dissertation.
In my dissertation I am proposing a cognitive linguistics exploration for poetic-rhetoric strategies used by Basque political poetry of the 70's, namely metaphor and almost idiomatic collocations. (My background includes having studied with professors as Mark Turner, J. Grady, Jeanne Fahnestock, and others at the University of Maryland).
I have also been teaching pedagogyc uses of corpora, especially related to second language acquisition Bayonne 2004). I am quite familiar with both TACT and Concordance softwares, for linguistic-semantic research purposes (see my work together with Iraide Ibarretxe).
But, at this point of my research I need something more stylistics centered, easy parsing mechanisms that could provide some rhetoric information about text-types and the like.

Name: Jonathon Gibson
English Higher Education Academy / Royal Holloway College

[HEA representative - no information given, but interested in teaching applications of electronic resources.]

Name: Neil Grindley
King's, London

[Methods Network representative]

Name: Geoff Hall
Senior Lecturer
CALS (Centre for Applied Language Studies)
University of Wales Swansea

1. Research
I am interested as part of ongoing work on the language and style of D. H. Lawrence to make some corpus linguistic investigations of his work and of literary critics' assertions about his style and language, but currently incompetent to do it myself. My knowledge of these fields is very much theoretical rather than hands on. I have read interesting looking work, and a long time ago did a Masters at Birmingham which inevitably interested me in the field, as well as (I hope) informing me about some of the basic issues and techniques, at least at that time. I'd be interested to identify some statistically (e.g. frequent) key terms in Lawrence's writings, 'mind', 'body', 'machine', 'organic' , 'nature', 'soul' or whatever they turn out to be, to see where and how they are used, perhaps to compare to other writers' uses. Another area it occurs to me where this kind of work might be interesting to pursue is in identifying clusters of lexis, patternings, repetitions, lemmas etc, his 'obsessive' style. A third possibility would be to look at revisions, as the work of the Cambridge editors has made so much
more of the progressive drafts of major works available. This probably all sounds too ambitious and unfocused, but that's why this kind of workshop would help me define better what to do and in what order or stages. I am also interested to learn what a corpus approach might teach as a 'discovery' approach rather than just testing the validity of pre-corpus critics' assertions.
I would value hands on practice to address some of these issues, as well as a better understanding of possibilities and limitations of corpus linguistics for such a project. I lack confidence and experience in this area.

2. Teaching
A secondary interest is definitely in terms of benefits for teaching and research supervision, where I now find myself often musing vaguely that 'a corpus investigation' might be useful here, but then hastily referring the student on to a colleague if they take me up on the idea.

Name: Ramesh Krishnamurthy
Email address: R.Krishnamurthy@aston.ac.uk
Institutional affiliation: Aston University
Where you heard about the event: a colleague forwarded Martin Wynne's email of c. 4th April

I have been involved in corpus analysis for lexicography since 1984 at Cobuild, University of Birmingham. I later played a major part in the construction of corpora for Cobuild, HarperCollins, the EU-funded TELRI project, Wolverhampton University, and have just started corpus development at Aston University. The focus here is on multilingual corpora (English, French, German, and Spanish) and parallel corpora for Translation Studies. We are also endeavouring to develop a pedagogically-oriented (as opposed to research-oriented) corpus software system. Many of my colleagues are involved in teaching literature and stylistics as well as language, and I would like to learn more about how corpus systems can be made more amenable to the investigation of literature and stylistics, and how the associated analytical approaches might differ from language-linguistic-focussed analyses.

Name: Lesley Kirk

I am starting a PhD at UCL's English department in September, having done their MA in English Language and helped to do some of the correcting of the Survey of English Usage corpus for the new release. My PhD will be on the development of Henry James' style using a corpus approach. I'm planning to tag and, I hope, parse at least 3 James novels and then look at a variety of features. At this stage I'm thinking about complexity, detached functions and the possible application of Biber's multi-dimensional approach. It's all very much at the thinking and reading stage, but some contact with other corpus work outside the Survey would be extremely useful. At the moment, apart from the Birmingham workshop and the Style in Fiction event, I've only really seen ICE-GB at work.

Although I'm not officially enrolled until September, I'm starting to read now, so the Oxford Workshop would be wonderfully timely in giving me a wider frame of reference, a variety of angles on corpus stylistics and perhaps some practical skills.

Name: Dr. Marina Lambrou
Email address: m.lambrou@uel.ac.uk
Institutional affiliation: University of East London
Where you heard about the event: via email from Workshop Organiser [at a PALA event]

I am the Programme Leader for Applied Linguistics and teach a module in Stylistics called Language in Literature. I am interested in how marked forms can be created by collocating pairs of words to create 'style' and unusual literary effects. I also teach a module called Analysing Language where I do a session specifically on collocations and lexical phrases. Here, I am interested in how meaning is derived from the way that words are collocated with each other rather than in isolation (-something that is pertinent to anyone interested in teaching EFL or studying translation).
I am looking forward to the two day workshop as it sounds extremely useful and will broaden my knowledge and understanding of the various methods of analysis of linguistic corpora. I particularly like the sound of the practical hands-on exercises with experts in the field.

Name: John McKenny
john.mckenny@unn.ac.uk <mailto:john.mckenny@unn.ac.uk>
English Language Centre, Northumbria University.
Hear about it from: CORPORA mailing list

I am finishing my Ph.D. in which I used a corpus-based approach to EAP phraseology and now wish to move on to do work in corpus-based stylistics. My thesis looked at naturalness and idiomaticity in non-native prose. I would now like to do a comparative corpus analysis of the literary works of certain non-native authors e.g Joseph Conrad and Liam O'Flaherty, using reference corpora of contemporary authors, to determine whether there are any distinctive features attributable to non-nativeness. As I will be teaching MA students with wide-ranging literary interests I would be pleased to learn some techniques and approaches which could assist in doing corpus-based literary or stylistic research.

Name: Dan McIntyre
Institution: Huddersfield University

I'd like to attend the workshop on Corpus Approaches to the Language of Literature. My research is primarily in literary stylistics and I would like to supplement my knowledge of corpus linguistics enough to be able to include corpus methodologies in my work in stylistics. I am currently planning a project with Michaela Mahlberg investigating authorial style and text-style in the work of Ian Fleming, so the workshop would clearly be relevant for that. In terms of disseminating the lessons and resources of the workshop, I am currently planning a third year undergraduate corpus linguistics module at Huddersfield that will be offered to both language and literature students (in effect, using stylistics as a bridge between language and literature), and I hope the workshop would help me in planning and running this course. I am also encouraging my PhD students to use corpus methodologies in their own research, and I hope the workshop would allow me to advise them more confidently on this. Finally, a number of my colleagues here who are stylisticians have expressed an interest not in corpus linguistics per se but in using corpus evidence as support in stylistic analysis. I would hope to be in a position to be able to help them with this aspect of their research.

Name: Miss Meng Ji
Email address: m.ji@imperial.ac.uk <mailto:m.ji@imperial.ac.uk>
Institutional affiliation: Translation Studies at Imperial College London
Where you heard about the event: Arts and Humanities Data Service

My current research at the Translation Centre of Imperial College London is a corpus-based study of the translations of Cervantes' Don Quixote into Chinese, which is an essentially unexplored area of linguistic studies in terms of the comparability of two genetically and geographically unrelated languages, namely early seventienth century Spanish and contemporary Chinese. It is very an interesting research topic and my current study entails the building-up of a parallel Spanish-Chinese corpus of D. Q. Thus I am particularly interested in acquiring more information related to the expertise of using corpus methods in textual analysis.

Name: Julie Millward
Email address: j.millward@sheffield.ac.uk
Institutional affiliation: The University of Sheffield
Heard of event: 'pala-announce' email circulation list

My main research area (PhD) is stylistics-based and concerns perceptual responses to invented language. In addition, I maintain an interest in textual representations of dialect, and have a developing interest in cognitive poetic approaches to neologism (more specifically, within Text World Theory). In short, all these are centrally concerned with non-standard language in literary texts. I have previously used concordancing software - WordSmith Tools - to locate examples of non-standard language, and would welcome the opportunity to further explore the potential of computational approaches to identify instantiations of non-standard language in texts.
Additionally, I am seeking a full-time academic post as my PHD nears completion, and would appreciate the opportunity to acquire useful transferable skills; in the meantime, I continue to undertake hourly-paid teaching in stylistics at the University of Sheffield, where the content of this workshop might productively contribute - to both learning and teaching - to the wide range of stylistics-based courses available there.

Name: Jacqui Mullender

I am currently a Master's student at the University of Birmingham, taking the Literary Linguistics MA there. I have begun studying corpus linguistics this year, with Nicholas Groom's MA module of the same name, but I am hoping to extend this in my MA dissertation this summer, analysing certain aspects of the lexis of Shakespeare's sonnets with WordsmithTools4 and the corpus of the poems. Although I have some basic grasp of corpus data analysis, (a rudimentary familiarity with Wordsmith Tools4, and a little experience using the Cobuild Bank of English) your workshop will be invaluable for me, increasing my understanding of the corpora and the software, and helping me refine and develop my search methods.

My PhD (which has been accepted for starting in Sept. 2006, and will be jointly supervised by the Shakespeare Institute and Professor Michael Toolan of the Birmingham English Dept.), will be an empirical analysis of the characterisation of Shakespeare's romance plays, involving not only the use of various models from speech act theory and discourse analysis, but also the identification of lexical saliencies which may be relevant in determining consistency or otherwise of characterisation and discursive focus, for which latter I shall need to work on a corpus of the four plays in question. I may well decide in addition that it is necessary to use further corpora to compare the lexis of these plays with the wider Shakespeare canon. Jonathan Culpeper's recent work on Shakespearean characterisation (for example his 2001 chapter 4) is very important to my proposed research, and I am keen to learn more from him.

Name: José L. Oncins-Martínez
Univ. Extremadura, Spain.

We've met a couple of times in PALA (N.York, Huddersfield...). I saw your workshop announced a few months ago but could not apply then for professional reasons (classes, departmental meetings...). Now I have managed to get a week off (May 15-21) and I think it would be a fantastic opportunity to attend the workshop and see you [Martin Wynne] and Jonathan again (we've had him for a Conference here a couple of weeks ago). My main reasons for wanting to enrole in the workshop is that at present we are engaged in a research project on Shakespeare and phraseology and I think this would be a good opportunity to catch up with methods of analysis with corpora.

Name: Dr Yanna Popova
University of Oxford
From July 2006: Case Western Reserve University
Cleveland, Ohio.

I heard about the event on the PALA list.

I've used some corpus techniques in my various stylistic explorations. For example, I've used the BNC corpus for synaesthetic adjective/noun combinations, limited metaphoric expressions and others. I would like to learn more as I find the approach extremely useful and potentially limitless in its scope for studying various kinds of texts. Some of the problems, however, need addressing and would like to raise some questions during the workshop.

Name: Gabriela Saldanha

I'm a lecturer in Translation Studies at Imperial College, London. I've heard about the workshop through the corpora list and I'm very interested to attend because of the nature of my research. My PhD looked at several literary translations from Spanish and Portuguese into English by two translators, and the aim was to identify stylistic traits that can be attributed to the translator rather than the author or to linguistic constraints imposed by the translation process. Apart from publishing in that area, I'm writing a book on corpus-based approaches to the study of translation to be published by Continumm. My aim for the book is to further develop corpus-based methodology within Translation Studies, bringing in insights and techniques from other areas, such as literary studies, stylometry, forensic linguistics, etc. I teach Translation Technology, Translation Theory and Corpus Linguistics at the MSc in Translation at Imperial College London.

Name: Violeta Sotirova
Email address: violeta.sotirova@nottingham.ac.uk
Institutional affiliation: University of Nottingham
Where you heard about the event: email from Martin Wynne [PALA list]

A short statement indicating your current or proposed areas of research and why you wish to attend this workshop: I work on the presentation of narrative viewpoint and some of my analyses have involved quantitative studies of the use of certain conversational features, such as sentence-initial connectives and repetiton. I would be interested to learn more about the use of software in collecting this kind of quantitative information about texts.

Name: Raksangob Wijitsopon
Email address: r.wijitsopon@lancaster.ac.uk
Institute: Lancaster University
Where you heard about the event: I got a forwarded email from Mick Short (my supervisor)

Why I wish to attend this workshop: My PhD research involves a corpus approach to style in Jane Austen's Emma. I'd like to investigate what stylistic choice are related to the theme of illusion-reality in the novel. Being a Lancaster student, I'm quite familiar with basic concepts in corpus work, e.g. keyword, collocation, cluster, etc. but I'm not quite sure how these concepts can be drawn on for work in stylistics.

Name: Xiaotian Guo (Mr)
Email address: xxg750@bham.ac.uk or garlickfred@gmail.com
Institutional affiliation: University of Birmingham
Where you heard about the event: Corpora mailing list

I am doing learner corpora, an approach called contrastive interlanguage analysis. Being near the completion of my PhD, I would like to attend this event because I wish to learn how corpus linguistic means could be used to expore literature works. I may join a research group who is analysing three versions of Hongloumeng, A Dream of Red Mansions, one of the three most well-known classics of Chinese literature.