One-day symposium
Corpus linguistics beyond boundaries
Interdisciplinary applications
Abstract booklet

Friday 10 July 2015
University of Nottingham
## Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00 - 09:15</td>
<td>Registration</td>
</tr>
<tr>
<td>09:15 - 09:30</td>
<td>Opening</td>
</tr>
<tr>
<td>09:30 - 10:00</td>
<td><strong>Sylvia Jaworska and Anupam Nanda (Reading University)</strong>&lt;br&gt;Doing well by talking good? Combining a corpus and computational approach with an econometric analysis to study the impact of Corporate Social Responsibility</td>
</tr>
<tr>
<td>10:00 - 10:30</td>
<td><strong>Camille Laporte (University of Leeds)</strong>&lt;br&gt;Adversarial strategies in political discourse: A corpus linguistics study of the 2012 U.S. election debates</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td><strong>David Wright (Nottingham Trent University)</strong>&lt;br&gt;Corpora in court? A case for a corpus-based approach to forensic authorship analysis</td>
</tr>
<tr>
<td>11:00 - 11:30</td>
<td>Coffee &amp; tea break</td>
</tr>
<tr>
<td>11:30 - 12:00</td>
<td><strong>Heather Froehlich (University of Strathclyde)</strong>&lt;br&gt;“How now, sir John?“: Locating social class in Early Modern drama</td>
</tr>
<tr>
<td>12:00 - 12:30</td>
<td><strong>Ivor Timmis (Leeds Beckett University)</strong>&lt;br&gt;Corpus linguistics and social history: A case study from the Bolton/Worktown corpus, 1973-1940</td>
</tr>
<tr>
<td>12:30 - 13:00</td>
<td><strong>Richard J. Whitt (University of Nottingham)</strong>&lt;br&gt;The linguistic realisation of evidentiality in Early Modern English medical writing</td>
</tr>
<tr>
<td>13:00 - 14:00</td>
<td>Lunch break &amp; posters display</td>
</tr>
<tr>
<td>14:00 - 14:30</td>
<td><strong>Nigel Slater (University of Nottingham)</strong>&lt;br&gt;The language of acute pain assessment: A corpus-based critical discourse analysis</td>
</tr>
<tr>
<td>14:30 - 15:00</td>
<td><strong>D’Egidio Angela (University of Salento)</strong>&lt;br&gt;Using Wmatrix to analyse the language of tourists</td>
</tr>
<tr>
<td>15:00 - 15:30</td>
<td><strong>Clyde Ancarno (King’s College London)</strong>&lt;br&gt;Corpus linguistics and anthropological data, which challenges and opportunities?</td>
</tr>
<tr>
<td>15:30 - 16:00</td>
<td>Coffee &amp; tea break</td>
</tr>
<tr>
<td>16:00 - 17:00</td>
<td><strong>Paul Baker (Lancaster University)</strong>&lt;br&gt;Big Data, big Impact? Corpus linguistics and press representation of Muslims</td>
</tr>
<tr>
<td>17:00 - 17:30</td>
<td>Closing</td>
</tr>
</tbody>
</table>
Corpus-based discourse analysis is an increasingly popular way of conducting research on representation of identity. Corpora, or large collections of representative texts, stored electronically (e.g. press, tweets, fiction, political speeches) are examined via computer software. Such tools summarise frequency information in large corpora and perform statistical tests in order to identify phenomena that are more frequent than expected. The tools can also present language data in ways that allow researchers to identify linguistic patterns more effectively. My own research has used corpus methods to examine representations of gay, bisexual and heterosexual men and women, trans people, refugees and asylum seekers and people receiving benefits. This talk describes an ESRC-funded project on the press representation of Muslims and Islam, focussing on methods used, key findings and the experiences of our team in relation to impact. In terms of academic impact, I will address challenges faced in trying to popularise a new method over the last decade, while in terms of impact outside academia I discuss obstacles to getting this kind of academic research acknowledged and accepted in public contexts.
ABSTRACTS

Doing well by talking good? Combining a corpus and computational approach with an econometric analysis to study the impact of Corporate Social Responsibility

Sylvia Jaworska and Anupam Nanda
Reading University

The aim of this paper is to demonstrate how the tools of Corpus and Computational Linguistics can be effectively applied to study language use in corporate disclosures and gauge its effects on financial performance. It will do so by examining a large corpus (14,915,714 tokens) of Corporate Social Responsibility (CSR) reports produced by 20 oil companies from 2000 to 2013.

Examining corporate disclosures, specifically CSR reports, is highly relevant because they are the most visible documents describing organisations’ actions and goals in relation to its stakeholders and society. There is evidence suggesting that corporate disclosures and their specific linguistic properties have an impact on company’s performance in that they influence market responses and investors’ decision-making. Despite the growing importance of CSR and the impact of corporate disclosures, there has been little research that examined the language of CSR reports.

The main questions which this research addresses are:

Q1: What are the key topics communicated in CSR reports?
Q2: How have they changed over time and in response to significant shock events (e.g. environmental disasters)?
Q3: Is there a relationship between the identified CSR topics and other financial performance indicators?

Topic modelling and collocational analysis are performed on the data to identify key messages of CSR reports (Q1) and their changes over time (Q2). These are subsequently correlated with selected performance indicators using simple econometric tests to see whether there is a relationship between the identified semantic categories and financial performance (Q3). In doing so, this study demonstrates how corpus-linguistic and computational methods can be used beyond the boundaries of Linguistics to provide empirical evidence for the interdependency between language and real-world issues, specifically issues pertaining to business and economy.
Adversarial strategies in political discourse: A corpus linguistics study of the 212 U.S. Election debates

Camille Laporte
University of Leeds

In this paper, I consider how adversarial relations are built in the context of live electoral debating between candidates, during the televised debate series which occurred in October 2012, in the midst of the United States (U.S) Presidential Election campaign of the same year.

This paper uses data from my purpose made corpus of 2.7 million words from political discourse in the UK, U.S and France. A multimodal discourse analysis of the debates (via the analysis of video extracts from the debate series) is carried out, in order to gain insights into the way in which adversarial discourse is conveyed via both verbal and non-verbal means in this confrontational context.

I discuss how questions and answers participate in the building of adversarial relations between two sets of participants: the candidates as well as the audience, whose voice is expressed by both the debate moderators and some audience questioners.

Drawing on Brown and Levinson’s theory of politeness (1987) and its application to political discourse (see Bull & Wells, 2012; Bull, 2013), I analyse how questions and answers participate in the creation of adversarial relations, while including rhetorical questions as a means of answering questions by the candidates. I also review how the presence of vague language can stimulate adversarial relations, and finally, I consider the various degrees of assertiveness and defensiveness in the candidates’ verbal communication.

Conclusions define the types of adversarial strategies used by political leaders, in relation to the political outcomes of each debate for the candidates, that is, how their respective performances were perceived by the public.
Corpora in court? A case for a corpus-based approach to forensic authorship analysis

David Wright
Nottingham Trent University

Forensic linguists are increasingly being asked to identify the authors of disputed or anonymous texts that are potentially evidential in criminal cases and to provide expert witness reports. Research into forensic authorship analysis, therefore, is largely focused on developing methodologies capable of producing evidence accurate and reliable enough to be permissible in a court of law. However, over the last decade, two distinct and seemingly incompatible approaches to authorship analysis have developed. On the one hand there are computational, statistical or ‘stylometric’ techniques, and on the other there are methods that are descriptive, qualitative or ‘stylistic’ in nature.

Recently there has been a debate in the field as to which approach is best equipped to meet the exacting standards required of expert witness testimony, both in the US and the UK. For example, some argue that a purely qualitative approach lacks the statistical reliability of computational methods, while others believe that the results produced using algorithmic ‘black box’ techniques are often impossible to interpret in linguistic terms and, in turn, are difficult to explain to lay judges and juries.

This paper outlines the main advantages and drawbacks of these respective approaches and, using examples from authorship work undertaken on a dataset of 2.5 million words of Enron emails, argues that corpus linguistics offers a means by which they can be effectively combined. Going forward, an appeal is made for researchers in authorship analysis to utilise corpus linguistic methodologies in such a way that statistical results can be underpinned by linguistic theory and supported by linguistic evidence.
"How now, sir John?": Locating social class in Early Modern drama

Heather Froehlich
University of Strathclyde

In this paper, I argue that social class can be identified and understood through the use of vocatives. Jonathan Culpeper and Dawn Archer (2003) claim that using nouns in the vocative mode is an unwise measure for identifying gentry-class figures in spoken discourse; however, as I demonstrate in this paper, new resources in the form of the Historical Thesaurus of the Oxford English Dictionary (HTOED; 2010) and the public-domain release of the first 25,000 hand-keyed and machine-readable texts from the Early English Books Online Text Creation Partnership (EEBO-TCP, phase I, 2015) make it possible to challenge this claim.

In the second part of this paper I ask if vocatives marking for class status correlate to character construction in early modern dramatic storyline. Character construction is a highly relevant process in dramatic storylines, in which individuals are introduced and information is gathered about them, which would then presumably drop off as characters become more established. Dramatic structure and narrative arcs are driven by storylines; Freytag’s (1863) pyramid is based on patterns he found in Shakespearean and Greek tragedy. [Reference] argues that issues of social mobility underlie the storylines of many early modern plays.

I use the HTOED to identify vocatives marking for social class and apply these to a sample corpus of Shakespearean drama. I then use a corpus of 332 Early Modern London plays curated from the EEBO-TCP initiative to explore issues of dramatic structure. Through a combination of concordance plots and close textual analysis, I examine whether vocatives can be used to identify characters of different social classes and test whether they have a role in character construction.
Corpus linguistics and social history: A case study from the Bolton/Worktown corpus, 1937-1940

Ivor Timmis
Leeds Beckett University

This paper is based on a corpus which has been drawn from the Worktown (Bolton) papers of the Mass-Observation archive, a sociological and anthropological study of the Bolton (UK) working class in the late 1930s. The corpus consists of snippets of conversations which took place in pubs, shops, streets and sports grounds in Bolton at the time, conversations which were transcribed 'live' by observers often operating incognito in the town. Compiling the corpus has involved identifying these snippets among the 400,000 pages of sociological records. The unusual genesis of the corpus offers, I argue, a good basis for interdisciplinary work which takes us beyond corpus linguistics into social history.

The case study I offer in this paper investigates the relationship between certain pronoun uses and identity among the working class of Bolton in the late 1930s. I argue that certain pronoun uses in the corpus can only be explained as homophoric reference, a kind of reference which depends on implicit agreement about the intended referent of the pronoun. This use is exemplified in the exchange below where the respondent (S2), a 55 year old working class woman from Bolton, assumes that her interlocutor will understand the referent of 'they':

S1: What do you think the country is fighting for?
S2: I don't know, they tell us what they're fighting for, don't they, it seems alright to me.

The article then discusses the basis on which this implicit agreement could operate: shared culture and knowledge and a tight network of social relations. I conclude that in this case the use a specific grammatical feature can only properly be understood in the light of the social and historical context.
The linguistic realisation of evidentiality in Early Modern English medical writing

Richard J. Whitt
University of Nottingham

This paper examines how sources of knowledge were conceptualised and linguistically expressed in Early Modern English medical writing. Evidentiality, the linguistic realisation of a speaker's (or writer's) evidence for an asserted proposition, has received significant attention in the last three decades (Boye 2012). However, only a handful of recent studies have considered the role of sociohistorical and cultural factors in the use evidential markers in specific genres of writing (Taavitsainen 2001, Grund 2012), and it is this gap which the current paper seeks to fill.

The Early Modern period witnessed the transition from scholastic and learned models of medicine, where classical authorities such as Hippocrates, Aristotle and Galen were considered infallible authorities on medical issues, to more empirical methods of medical knowledge and scientific inquiry (Wear 2000; Lindemann 2010). Consequently, it is predicted that there will be a decrease in the use of markers of mediated information (i.e. citing classical authorities such as Galen and Aristotle) and an increase in the use of markers of direct observation and inference. Data from the Corpus of Early Modern English Medical Texts (Taavitsainen & Pahta 2010) will be analysed using a combination of “bottom-up” analysis (close reading of selected texts [see Bednarek 2006, Grund 2012]) and subsequent “top down” searches for frequently used evidential markers using the WordSmith concordancer program (Scott 2012). The focus here will be specifically on medical treatises and textbooks covering issues ranging from the plague and other diseases to midwifery to medical diagnoses to therapeutic substances (Pahta & Ratia 2010; Taavitsaien & Tyrvkkö 2010), thus illustrating how corpus linguistic analysis can assist us in tracing the changing epistemologies of medical culture.
The language of acute pain assessment: A corpus-based critical discourse analysis

Nigel Slater
University of Nottingham

“Do you have any pain?” may be seen as a good question to ask when assessing pain. However, it is not necessarily an easy question to answer. Whether in a post-surgery context in an acute hospital or in other healthcare settings, what that pain is, how it manifests itself, and what any patient has to say about their pain could be very different. This complexity in the pain assessment process and some of the difficulties this may present for clinical staff and patients warrants detailed examination. In a response to a call for more detailed understanding of pain assessment this paper concentrates on the complexity of language by examining the recordings of patient health care in naturally occurring interactions during pain assessments and analysed through a corpus linguistics based critical discourse analysis approach. This is something that has not been achieved before in an acute hospital ward setting.

Two key areas were identified, the first related to the traditional aspects of pain assessment: the terminology used, location and function of pain. The second more important area related to how healthcare professionals presented a certain ‘mentality’ about the assessment process in how they appeared to be patient centred but through the use of brevity of interaction and trivialisation of the issues actually presented an opposite view.

The primary conclusion is that although healthcare workers apply pain assessment processes, their use of language can show that they are both patient-centred and have their own motivations and agendas.
Researchers in corpus linguistics have to date investigated tourism discourse through methodologies such as collocation, wordlist and keyword analysis, focussing on cultural representations, stylistic choices and linguistic strategies (e.g. Bianchi 2012; Cappelli 2012; 2013a; 2013b; D’Egidio 2009; 2014; Fina 2011; Francesconi 2007; 2012; Maci 2013; Manca 2011; 2013; Nigro 2006). With the exception of Maci (2013), several scholars inductively and manually classified findings in semantic categories.

This paper explores the effectiveness of the use of WMatrix (Rayson 2009) as a first step for analysing the language used by tourists. This web-based corpus analysis and comparison system was chosen as a means to automatically assign semantic tags to two comparable corpora of travel blogs and trip reports written in English about two tourist destinations in Italy: the popular city of Florence, and Puglia, an off-the-beaten-track destination in the South of Italy. The USAS tagger was useful for identifying the most frequently occurring semantic categories. The automatic semantic analysis of the corpora and the extraction of the most frequent semantic fields, along with a careful examination of the dominant lexical items within them with WMatrix, and subsequent identification of patterns in concordances of lexical items by means of WordSmith Tools 6.0 (Scott 2012), proved to yield comprehensive results for the analysis of the tourist gaze (Urry 2002; Urry and Larsen 2011).

This methodological approach was particularly helpful in identifying a number of similarities and differences in the way tourists writing in English experienced, perceived and evaluated the two Italian tourist destinations.

The results also demonstrated that traveller-generated content can provide invaluable feedback for the tourism industry. Translators can have an idea about how to re-mediate tourist materials in order to produce more effective translated tourist texts into English as a Lingua Franca, which meet tourists’ expectations.
Corpus linguistics and anthropological data, which challenges and opportunities?

Clyde Ancarno
King’s College London

Corpus linguistic techniques are used by an increasingly diverse range of social scientists to support the analysis of their text-based data. The data I focus on in this paper is questionnaire data gathered as part of an anthropological project: ‘Knowing each other: Everyday religious encounters, social identities and tolerance in southwest Nigeria’. The latter examines the coexistence of Islam, Christianity and traditional practice in the Yoruba-speaking parts of southwest Nigeria, i.e. how our participants encounter other religions and how their identity is shaped by these interreligious encounters. The questionnaire data was collected in southwest parts of Nigeria in 2012-13 and involved more than 3000 participants.

The aim of the paper is threefold. I will explore the nature of our cross-/inter-disciplinary collaboration. I will examine the challenges we faced, e.g. using corpus tools to explore data which was not devised with corpus linguistics in mind. I will also allude to some of the challenges related to the fact that our corpus is bilingual (English and Yoruba), giving preliminary insights into the limitations of using corpus linguistics tools to engage with the Yoruba data in particular. Third, I will highlight how our ongoing work has led us to focus our attention on the extent to which our methodology allows us to (i) gain general insights into the very topic of our data, religious encounters in multireligious communities and to (ii) draw a detailed picture of the negotiation of religious identities among Muslims and Christians in Yorubaland.
Extraction and analysis of modal auxiliaries in consecutive clauses from the BNC
Robert Chartrand
Kurume University, University of Birmingham

Historical corpus stylistics: Keywords and moral didacticism in Sir Gawain
Jacqueline Cordell
University of Nottingham

Manipulation and persuasion in political discourse: A corpus linguistic and critical discourse analysis of Barack Obama and George W. Bush’s political speeches
Tara Coltman-Patel
Nottingham Trent University

The profound shoom of lip-music: Corpus-based analysis of an invented literary language
Jim Clarke and Benet Vincent
Coventry University

Media use of pharmaceutical promotional literature in Britain and China: The case of antidepressant medication
Fang Wang
University of Birmingham