Topics in Corpus Linguistics for Social Media Research

Organiser: Amanda Potts, Lancaster University

Half-day workshop

Social media platforms are at the cutting edge of technological advances, and research in discourse on these sites is likewise leading the way in a range of interdisciplinary investigations. Discourse drawn from social media is a natural fit for corpus linguistic studies: it is plentiful, digital, and natural. However, the emerging field of corpus-based social media studies is encountering issues inherited from its various parent disciplines as well as facing growing pains unique to the sub-discipline.

In this half-day colloquium, a range of topics in corpus linguistics for social media research will be presented by invited speakers, with ample time given to discussion of emerging issues in participants' projects. Talks move through an evolution, with Dawn Knight and Svenja Adolphs opening the day by discussing where e-language falls on the spoken/written continuum, and how we can make sense of social media discourse as a genre. Next, Chris Pak and Alison Sealey demonstrate their method of identifying and filtering tweets for a large-scale critical discourse analysis of discursive ecologies.
Rachelle Vessey takes another view of Twitter, underscoring the superdiverse nature of social media data by drawing upon examples of bilingual tweets. Finally, Amanda Potts takes a triangulated approach by considering the production of YouTube videos by very popular social media users, and the impact that reception of these has on the discourse of their fan communities. Extended descriptions of these talks appear below:

1. **Language in the Digital Age: Revisiting the Speech-Writing Continuum**
   Dawn Knight, Cardiff University
   Svenja Adolphs, University of Nottingham

   As a relatively new 'genre' of communication (Herring 2002), the definition and description of the features of e-language (defined here as any communicative, interactive and/or linguistic stimulus that is digitally based and 'incorporates multiple forms of media bridging the physical and digital' (Boyd and Heer 2006: 1) and how it compares and contrasts with spoken and written genres of communication is an on-going concern in studies of Computer Mediated Communication, Applied Linguistics, Corpus Linguistics and beyond.

   Early work from Crystal (2003: 17), suggested that spoken and written language effectively exist on a 'continuum' of formality (also see Condon and Cech 1996; Ko 1996; Herring 2007 for further discussions on the differences between spoken and written discourse). The 'more' formal language structures exist on the left of the continuum, where written language is conventionally positioned, and the least formal exists towards the right end of the continuum, where spoken language is conventionally perceived to be positioned. So where does e-language fit on this continuum? And is it more 'spoken' then 'written' when we look at the characteristics of language used, beyond this notion of formality?

   The current presentation explores these questions, by providing a corpus based investigation of the similarities and differences of the use of e-language (as evidenced by analyses of the Cambridge and Nottingham E-Language Corpus, CANELC) in comparison to spoken and written records of communication (as evidenced by the British National Corpus, BNC). The paper begins with a discussion of the most frequent words/phrases used in CANELC (within and the across different modes) in terms of word *function*, *sense* and *meaning* and how these compare to the spoken/written elements; it also comments on the incidence and frequency of modal verb usage and hedging devices in e-language and then questions whether e-language appears more or less (in)direct and/or (im)polite than spoken and written discourse.

2. **The Discursive Representation of Foxes and Bees in the Twittersphere**
   Chris Pak, Lancaster University

Alison Sealey, Lancaster University

As one strand of a funded research project that explores how people talk and write about animals (http://animaldiscourse.wordpress.com/), we are constructing a corpus comprising texts from a wide range of genres. One of our sub-corpora consists of tweets, and we have had to find a method of identifying tweets that are about various kinds of animals, while eliminating those where terms denoting animals are used metaphorically (e.g. 'yeah baby i am an ANIMAL in bed more specifically a koala I can sleep for 22 hours a day'). This presentation will discuss the methods we used to identify relevant tweets and process them for inclusion in our corpus. It will explain some of the approaches we have been using to capture relevant data, including filtering handles and hashtags, as well as tools such as Topsy, which sifts and retrieves tweets published since Twitter's founding in 2006, and Keyhole, which is a social monitoring database that generates statistical summaries of the circulation and distribution of tweets.

The talk will include analysis of some of the ways foxes and bees have been tweeted about. These two animals have been selected as a way of exploring the networks of associated animal terms or "discursive ecologies" that appear in the corpus. They are associated with the British countryside, but are increasingly becoming urbanised. Central questions to be addressed in the presentation include whether Twitter offers a platform that allows users to give animals a political voice or to speak on their behalf, and whether Twitter offers users a space to develop a critical voice to assess news media and events about animals. Other questions considered are how these tweets assess and construct power relations between humans and animals and how far the referentiality of tweets allows for power to be interrogated.

3. Corpus linguistics and superdiversity

Rachelle Vessey, Newcastle University

"Superdiversity" has recently become a buzz-word in social research, signifying the unprecedented intersections of diverse factors that become meaningful within a "post-multicultural" world of migration and transnational flow (Vertovec, 2007). In linguistics, superdiversity has become an important new framework for understanding the complex ways in which new and changing meanings are attributed to languages (e.g. fluency, literacy, and multilingualism) and linguistic features (e.g. accent, code-switching, spelling) (Blommaert & Rampton 2011). Social media are significant site of information flow, relying on new and creative methods of communication. Although it is not yet known precisely how linguistic resources (including multilingualism) are deployed in these media spaces, it is clear that multilingualism figures to a large degree, even if English continues to be a powerful lingua franca. Accordingly, social media and the language used on social media sites have been argued to be "superdiverse" (Leppänen.and Häkkinen, 2012).

This paper explores the extent to which corpus linguistic frameworks can be used to explore superdiversity. Although some corpus linguistic research addresses multilingualism (e.g. through parallel, learner, and contrastive corpora), the field is dominated by corpora of more-or-less discrete languages (e.g. English, French, Spanish) and language varieties (e.g. British, American, and Canadian English). Thus, there are fewer multilingual or "superdiverse" datasets (Seidlhofer, 2012), and the field of corpus linguistics has not yet adopted or applied notions of superdiversity, despite the alignment of corpus linguistics and sociolinguistics (e.g. Baker 2012) and the burgeoning literature on superdiversity and sociolinguistics. Moreover, the evolving norms in online spaces mean that frequency and collocation do not have clear-cut meanings, especially on sites with specific constraints or affordances (e.g. character limitations, automated spell checkers, sharing or "retweeting"). As such, there are clear challenges involved in combining corpus linguistics and superdiversity, which this paper demonstrates by drawing on three separate case studies of Twitter data.

4. Considering YouTube Fandoms as Pro-Social Communities of Practice

Amanda Potts, Lancaster University

In the digital age, users of social media have the opportunity to craft the image of their identity presented to the larger public; 'when we step through the screen into virtual communities, we reconstruct our identities on the other side of the looking glass' (Turkle, 1995, p. 177). Scholarship on identity construction in online games and other simulated places is on the rise, as these are sites where identities are negotiated and (re-) constructed (or mixed and matched). Multiplayer online video games are a form of social networking which expose unprecedented numbers of gamers to scenarios, discourses, and identities far outside of their usual environments. In the case of gamers who broadcast their gameplay online, such as on the popular video sharing website YouTube, this also has the effect of impacting identity construction and perception of viewers and fan communities. In many ways, the internet is 'changing the way we think, the nature of our sexuality, the form of our communities, our very identities' (Turkle, 2004, p. 19).

This paper explores queer discourses produced by a group of very popular professional video game players on social media, with particular focus on the impact that this has on the language and interactions of the fan community. Three sources of data have been used: 63 YouTube videos, a corpus of 217,916 comments on these videos, and a 40-minute interview with a gamer featured in these videos.

Uniquely, though the majority of the discursive data is drawn from interactions between heterosexual males, introduction of homosocial and homosexual innuendo into the 'canon' (or source text) has granted a unique opportunity for very large adolescent fan audiences to encounter, interpret, and experiment with queer discourse (Tosenberger, 2008) in the comments sections of these videos. I demonstrate how, in this fan community, homophobic expressions have become stigmatized in their most derogatory forms or gained ironic meaning, thereby building solidarity and allowing for creativity. Incorporation of queer and nonheteronormative discourses in these videos—watched by hundreds of thousands, subscribed to by over a million—has a nearly unprecedented opportunity to undergo a trickle-down effect on viewers, who are involved in the participatory culture of the social media platform. Commenters have been observed to use the online community created by the fandom to explore their understanding of homosocial and homosexual relationships, "'working through social experiences and concerns'' (Jenkins, 1992, p. 215) by voicing their own emotions and experimenting with establishing new norms.

By demonstrating some key considerations and challenges encountered in a range of social media research, we hope contribute to a wider conversation on the many ways in which corpus linguistic methods may be integral to this type of work moving forward. We very much look forward to participation from colloquium attendees, including any insights that they may offer from their own work in similar or related fields. All are welcome to attend.