CANELC – Constructing an e-language corpus

Dawn Knight[1], Svenja Adolphs[2] and Ronald Carter[2]

This paper reports on the construction of CANELC: the Cambridge and Nottingham e-language Corpus[3]. CANELC is a one million word corpus of digital communication in English, taken from online discussion boards, blogs, tweets, emails and SMS messages. The paper outlines the approaches used when planning the corpus: obtaining consent; collecting the data and compiling the corpus database.

This is followed by a detailed analysis of some of the patterns of language used in the corpus. The analysis includes a discussion of the key words and phrases used as well as the common themes and semantic associations connected with the data. These discussions form the basis of an investigation of how e-language operates in both similar and different ways to spoken and written records of communication (as evidenced by the BNC - British National Corpus).

Keywords: Blogs, Tweets, SMS, Discussion Boards, e-language, Corpus Linguistics

1. Introduction

Communication in the digital age is a complex many faceted process involving the production and reception of linguistic stimuli across a multitude of platforms and media types (see Boyd and Heer, 2006:1). While a wealth of corpus research has been carried out on individual forms of e-language (i.e. language communicated via any digital resource), from SMS messages, to blogs and emails, corpora utilised to date tend to be either small scale or bespoke, that is, planned or utilised to answer very specific linguistic enquiries, and/or they consist of only one e-language variety (see Schler et al., 2006, Tagg, 2009 and Puschmann, 2009 for examples).

The most notable, large-scale selection of examples of current e-language corpora are detailed in figure 1. While invaluable for examining language patterns of their own individual text type/ language variety, such corpora are limited in utility. Although it is widely acknowledged that we live and communicate in a ubiquitous, digital world, the ways in which we actually do this, across multiple resources, remains an underexplored area of research in corpus linguistics as there is a lack of appropriate resources in existence to enable us to do this.

The next phase of corpora development should therefore seek to fill this void and integrate a wider range of different and relevant digital resources in a large-scale functional database for linguistic analysis. Arguably, a seemingly logical way of developing an integrated e-language corpus would be to attempt to combine all current corpora and thus build a corpus of already existing legacy data. While in theory this would allow us to construct a large scale corpus in very little time, with, it is assumed, minimal effort, in reality various practical and ethical challenges would be encountered when trying to do this.

Name	Description	Reference
Blog Authorship corpus	Freely available 140 million word English blog corpus. Comprises 681,288 blog entries taken from 19,320 bloggers over three different age groups.	Schler et al., 2006
CorText	110,000 word corpus of SMS messages (19,000 messages from 235 users), also with associated biographical metadata.	Tagg, 2009
Dortmund Chat-Korpus	1 million tokens from online chatrooms in German (140,000 chat conversations).	Beißwenger, 2007
Enron corpus	70 million words from emails sent by 150 individuals, mainly senior managers of the ENRON firm.	Klimt and Yang, 2004
Junk Email Corpus	373,000 words from 1563 junk email messages.	Orasan and Krishnamurthy, 2002
NPS Chat Corpus	10,567 messages from online chatrooms in English.	Forsyth and Martell, 2007
Twitter_Smallcorp	2 million word corpus of tweets.	Puschmann, 2009
ICWSM conference (TREC Tweets2011)	16 million word corpus of tweets.	Horn et al., 2011

Figure 1: A selection of current e-language corpora.

First, each of the corpora have likely been constructed in a different way, using different methods for extracting and storing data, with different header and related information being retained in each of them. Furthermore, the Blog Authorship corpus, for example, is already a few years old so the content is relatively outdated. This may limit the extent to which we can use analyses from this corpus to discuss patterns of e-language use in the 2010s and it also limits the comparability to the other e-language corpora in existence, as they all contain data from different time periods.

In order to ensure that the content of the corpus is structurally and compositionally consistent, accurate and as up to date as possible, it was deemed more viable to construct a new e-language corpus from the bottom up, one that preferably includes data from the past one to two years. The remainder of this paper introduces one such corpus, the newly constructed CANELC (Cambridge and Nottingham e-Language Corpus) corpus. The paper outlines the basic composition of the corpus, the approaches used in recruiting contributors and compiling the data before providing some results from preliminary analyses of the data, focussing specifically on defining some patterns in function, sense and meaning used in the corpus and how these compare with spoken and written components of the BNC (British National Corpus).

2. Why build an e-language corpus?

The motivation for building CANELC was two-fold. Firstly, from the perspective of Cambridge University Press, it was conceived as a potentially invaluable teaching and learning resource. It has been designed therefore with the purpose of informing and supporting content included in text books

and grammars published by the Press, with specific extracts of the corpus to be used to illustrate ideas and/or specific functions and properties of language usage.

Our own academic interest in CANELC encompasses a broader range of concerns, from not only providing the facilities for exploring patterns of lexical, grammatical and semantic properties of language use within and across different communicative modes, but also to helping us to develop our understanding of how these patterns of usage compare and contrast to those seen in previous corpus-based studies of spoken and/or written discourse. Basically, we are aware that a tweet or a thread on a discussion board, for example, is lexically and structurally different from standard written and spoken English, but exactly *how* and *why* they are different (and in which ways: from each different text type to the next) are questions that have yet to be fully explored.

Analyses of the CANCODE spoken corpus (Cambridge and Nottingham Corpus of Discourse in English[4]), in comparison to written counterparts from the CEC (Cambridge English Corpus[5]), for example, indicated that spoken language features a marked increase in the use in personal pronouns, discourse markers and response tokens in comparison to written language (see Carter and McCarthy, 2006: 9-16). So words such as *it's*, *yeah*, *I*, *you*, *you know*, *mmm* are indicative of spoken discourse and are considered to be markers of interactive informality. More formal linguistic structures such as *whom* and *no one* (in comparison to 'nobody, as is often used in spoken language), were found to be comparatively more frequent in the written components of CEC.

Crystal (2003: 17) suggests that spoken and written language effectively exist on a continuum of formality. The more formal language structures and conventions exist at the written end of the

spectrum and the least formal exists towards the spoken end. The question is where forms of e-language exist on this continuum. Crystal suggests it is perhaps somewhere in the middle, as a distinct form of language in itself, but where exactly this lies is still under debate by researchers in the field. Analyses of data from CANELC will hopefully allow us to make better informed judgements about the nature and characteristics of e-language and its 'best fit' along the continuum, providing the foundations for better enhancing our descriptions and understanding of language use in the modern digital age.

The issue of levels of formality in specific types of e-language has already received attention from researchers (see works by Sutherland, 2002; Shortis, 2007; Crystal, 2008; Hard af Segersteg, 2002 for further details). Tagg (2009) and Ling (2003) both report on the tendency for most SMS messages to be immediate and personal, written in the first person and directed to specific recipients. Tagg adds to this, suggesting that 'the informal and intimate nature of texting encourages the use of speech-like language' (2009: 17, also see Oksman and Turtianen, 2004). Similarly, Baron argues that email, as with texting and other common forms of e-language, is a written mode of communication but 'participants exploit it for typically spoken purposes' (1998: 36); it perhaps shares therefore more similarities with communication situated at the spoken rather than written end of the continuum. The blurring of traditional characteristics of spoken and written language in digital communication is something that has been discussed at length, although there is a limit to which this has been supported by corpus-based analysis of real-life data (see Biber, 1993; Collot and Belmore, 1996 and Crystal, 2001 for further details). The CANELC corpus enables such investigation.

The level of formality in text is closely tied to the function of the message, which poses a variety of

challenges when classifying text types, as non-typical features may be included in a text, perhaps as an expression of creativity or style or because the *context* in which they are used causes these changes in language use (for discussions of language and context see Labov, 1972; Bates, 1976; Nelson et al., 1985; Brown, 1989; Halliday and Hasan, 1989; Duranti and Goodwin, 1992; Widdowson, 1998; Green, 2002 and Scollon and Scollon, 2003). Who sends a message to whom and when and where this occurs can impact on the meaning and the pragmatic function of the content. SMS messages sent from a business director to a managing director of a different company, for example, are likely to be more formal than a message between two friends arranging a coffee date. Given this, information such as the date and time messages were sent and the identity of senders and recipients, as relevant (including age, gender, occupation, nationality and relationship), should be retained as metadata information when constructing new corpora of this nature. This information can then be consulted when analysing the data in order to help reconstruct elements of the fragmented context of the language-in-use and to help explain why certain patterns may exist.

It is unlikely that complete metadata records of e-language contributors can ever be constructed, as users often engage in a certain level of anonymity when online (especially when sharing data) and engaging in forms of e-communication. Furthermore, the notion of 'context' is in itself difficult to define and qualify, as is the extent to which it shapes or develops the meaning of a message. This is because context is a complex, fluid notion that involves social, physical and temporal dimensions which are often abstract. For example, a location may be defined by the use of the absolute; a specific grid reference on a map, street X. To an individual stood in street X, perhaps sending an email or writing a text, street X may be the location of a public house or a pool hall, a place where the contributor visits with specific friends or colleagues, meeting at certain times of the week to

partake in a particular activity, for example. Understanding and accounting for these more complex structures of the social context of the message will allow us to enhance our descriptions of language in use, providing an insight to more pragmatic, functional aspects of communication.

In practice, however, it is unlikely that such enriched information can be successfully gathered when constructing large-scale corpora, as such detailed enriched profiling of language is only really practical on a small-scale with limited numbers of contributors involved. Despite this, it is still important that we at least aim to collect some basic forms of metadata, including biographical information about contributors, where they are in the world, and categorising the intended readership of content, as this may still prove to be of interest when examining the corpora in more detail.

3. Composition of the Corpus

3.1. Data included

There are a range of different e-language resources that are used as a means of communicating in everyday life: from SMS messages to email activity, blogs, status updates on social networking sites and instant message conversations. CANELC comprises the data listed in Figure 2[6].

As outlined by Herring, there are a variety of different ways of classifying computer-mediated discourse. For the purpose of CANELC, the data is broadly categorised under a range of different 'genres' (Herring, 2002) with the overarching grouping of 'e-language'. These genres are essentially individual 'socio-technical modes', each of which is likely to have specific 'social and cultural practices that have arisen around their use' (Herring, 2007: 3). Coupled with the addition of

metadata detailing not only the specific mode of language, but also information of the 'participant characteristics', 'topic or theme' and so on (Herring, 2002: 19 - see sections 5 and 7), this broad method of categorization provides a way-in to exploring patterns of language use and to carrying out corpus-based linguistic research at the communicative *mode* level.

Data Type	**Number of Contributors**	**Number of Messages/ Entries**	**Word Count**	
			Raw	**%**
Twitter	30	18972	259101	26%
Blogs	36	1101	267983	27%
Discussion Boards	12	2715	242727	24%
Emails	Various	1920	128951	13%
SMS	11	5215	101913	10%
		29923	**1000675**	**100 %**

Figure 2: The contents of CANELC

CANELC was also constructed to allow for the querying the data, at a more general level, of the *genre* of communication (that is 'e'-based language or 'netspeak' - Crystal, 2001). This is because,

despite being *different* socio-technical modes, there is a key similarity between them insofar as they are all forms of asynchronous communication systems. Although SMS and email constitute interpersonal communication exchanged between a potentially large but bounded number of participants (discussed further in section 4), and Twitter and blogs are instead usually publicly accessible, none of these different modes 'require that users be logged on at the same time in order to send and receive messages' (Herring, 2007: 13). Instead, as with most written forms of language, these 'messages are stored at the addressee's site until they can be read' (Herring, 2007: 13). This is different to spoken language which is, conversely, most often synchronous.

Herring underlines that this makes 'synchronicity a useful dimension for comparing different types of CMC [computer mediated communication] with spoken and written discourse' (Herring, 2007: 9 - also see Condon and Cech 1996 and Ko 1996) and, although specific differences in patterns across the individual 'modes' are underlined, it is this dimension that motivates the preliminary comparisons between spoken, written and 'e-language' carried out in the final part of this paper.

3.2. Recruitment and permissions

To collect data for CANELC, authors of 'popular' blogs, discussion boards and tweets were targeted as it was thought that this would best represent the types of the discourse that the general public would be reading. This notion of 'popularity' was gauged according to the following requisites:

- Sites had to feature within online directories[7] of the most popular blog/tweet lists (sourced by Googling 'top ten blogs', 'popular blogs' and so on).
- Tweeters were to have at least 1000 followers.

- Posts from blogging sites had either a range of readers/followers and/or numerous responses to posts, indicating a large readership.

These were sites with ‘public’ rather than private profiles. From these lists, the specific individuals contacted (as hundreds of individuals were included here) were chosen at random in the first instance and were then filtered further by means of checking whether the following criteria were met:

- The prospective site was managed by a single individual (to ease problems associated with permissions for multiple contributors), who assumed copyright for their own material.
- Email/contact details were easily obtainable.
- The site contained a reasonable amount of text.

3.3. Gaining permission

Hundreds of potential target sites were shortlisted using this approach. Contact details of the owners/moderators of the sites were tabulated, with each being contacted to ask for permission to use their data. The permission process was tested during the piloting phase. The initial approach was to send a traditional consent form attached to an email detailing the aims and objectives of the study, then requesting each individual to firstly provide permission in a response to the email, then to sign the form and return it to the researcher.

30 prospective blog and twitter contributors were contacted during this piloting phase and while 12 individuals responded, only 7 of these provided full permissions. 5 others declined to participate and the remainder neglected to respond. The positive response from the 12 individuals was reassuring

but it was decided that a more streamlined approach for providing consent was needed as the process of posting and/or scanning in a long and detailed form was time consuming and inconvenient. As a second parse, instructions regarding the provision of consent were written into the initial correspondence sent to prospective contributors, making the process more streamlined. This allowed individuals to simply respond with 'yes, I provide consent', without them having to go through the more laborious form signing process.

Striving for consistency in the type of correspondence and documentation sent to each prospective contributor was of paramount importance to this project given that extracts of the corpus are likely to be published in academic texts and teaching materials. Therefore, an email and permission form was drawn up in consultation with CUP and their legal team in order to verify the legitimacy of the permission sought. This was circulated to over one hundred potential sites/individuals and in instances where 'full' permission was granted data was sampled. Permissions were not sought, and data was not taken from third parties who, for example, responded to content on a blog. However, a note of how many responses were associated with specific contributions *was* made in order to enrich the dataset.

With the discussion board data, consent was requested in the same way as already discussed, but as an additional measure discussion board moderators were asked with whom the sole copyright of content lay. If it was with the moderators themselves, content was taken from *all* users adding to the board. If not, individual contributors *as well as* the moderator were contacted and asked for permission to use their text. Again, only text provided by fully consenting moderators and/or individuals was used in CANELC.

3.4. Profile of contributors

CANELC aimed to include contributions from a range of different sociolinguistically profiled participants (that is, of different ages, genders and so on). As far as possible requisites identified in the 'aimed composition' column of figure 3 were to be met to ensure balance and consistency in the data. The 'actual composition' column of this table describes the extent to which these were met.

3.5. Access

Initial plans were to make this corpus open access and usable by all. Unfortunately, ownership and distribution rites enforced by our partners have resulted in the corpus being restricted in access. It is thus only available to researchers at the University of Nottingham or staff working at the Press.

Variable	**Aimed composition**	**Actual composition**
Number of participants	10 – 40 per source	11 – 36 contributors per source
Gender	50:50 male and female	50% of the corpus has a circa 50:50 balance. For 50% genders are unknown.
Age	Under 19 –	Contributors were from a range of different age groups

	10% of all data 20-24 – 10% of all data 25-29 – 10% of all data 30-34 – 10% of all data 35-39 – 10% of all data 40-44 – 10% of all data 45-49 – 10% of all	although the most populous groupings were 20-24 and 25-29 (there was not a strict balance of contributions across the groupings).

	data 50-54 – 10% of all data 55-59 – 10% of all data 60+ – 10% of all data	
Time period	Contributions posted from 2006-2011	Data from each contributors was collected over a minimum of 3 days, the majority within the 2010-2011 period.
Location	100% posting to sites ending in .co.uk	All sites ended in .co.uk and most contributors identified themselves as being British.

Figure 3: Profile of the contributors to CANELC.

4. Data types[8]

4.1. Tweets

It is estimated that over 175 million people use twitter (see www.twitter.com) globally, to update their 'followers', friends, and/or the world at large on their thoughts, feelings and reflections at a given moment. It is often used in a professional capacity, for publicity or advertising, but is also used on a more personal level, for individual tweeters to talk about their daily lives. Twitter operates in a similar way to Facebook (see www.facebook.com) updates and SMS messages in that it is restricted in terms of the number of characters (140) that can be inputted on a Tweet at any one time. But a 'tweeter' is able to contribute a potentially infinite number of messages over the course of a day.

An increasing number of linguistic studies have been carried out on the language of tweets (for examples see Borau et al., 2009; Honeycutt and Herring, 2009; Jansen et al., 2009 and Zappavingna, 2011) and, as identified in the introduction, there is an increase in interest in building and using twitter corpora, particularly in the field of Natural Language Processing (NLP), for the purpose of sentiment analysis (for examples see: http://www.tweetfeel.com/, www.sentiment140.com, http://tweetsentiments.com and http://www.tweettone.com/).

Tweets, in the same way as our second data 'type', blogs, can be classified as 'outward facing' forms of digitally based communication, in so far as they are posted on sites which can be accessed by anyone (unless they are hosted on members only sites) and so, it can be assumed, are aimed at a wider readership and audience than a personal SMS or IM (Instant Message - another form of e-language). The readership is often less specific although the content of the material may be

of interest to some individuals more than others. For example, a middle aged university lecturer may be more interested in the content posted by a publishing house, research network or fellow academics, rather than that posted by Britney Spears or the pop group JLS.

A challenge was posed when trying to decide which tweeters to target when constructing the CANELC corpus. We wished to collect data which was as 'representative' of each different e-language type as possible rather than simply use a web-crawler or API to collect data, randomly selecting sources. To achieve this we decided to collect data from popular public sites only (see section 3.2), ones that discuss a range of different topics, have as large a readership as possible and whose authors provided full permission to reuse their data. The selection and classification of topics was consistent to the approach used when collecting blog and discussion board data, as defined in section 5.

4.2. Blogs

While the use of Tweets has a relatively short history, with Twitter only being launched in 2006, the use of weblogs (blogs) first saw a 'sudden rise in prominence in 1999' (Myers, 2010: 10) and they are now authored by billions of web users across the globe. Blogs are generally longer excerpts of prose as they are not restricted by space or word count, so can run from a few sentences to numerous paragraphs of text. 'Blogging software means that anyone with access to the internet can post their thoughts, links and photos on a blog' (Myers, 2010: 77) although the readership of a given blog is again dependent on who is writing it, the topics covered by the content, accessibility to the content and how the blogs are presented.

There has been an increasing amount of research being carried out on blogs in the past area. One key area of study has focused on exploring patterns of language use and the social functions of blogging (see Gillmor, 2004; Allan, 2006 and Myers, 2010). The inclusion of blogs in CANELC aims to complement this already existing research. It also aims to allow us to examine the relationship between this mode of e-language and other varieties.

4.3. Discussion boards

Discussion boards are more interactive spaces for online communication. In a similar way to IMs and interaction on social networking sites (SNSs), individuals add comments around a given topic, either prescribed by the site moderator or by the first contributor to a thread and others read and respond to the comment, supporting, challenging and/or building on what has been said. Research on the social dynamics of internet forums has been widely published, often exploring the notion of the generation of a 'virtual community' through language (Jones, 1997) in a 'virtual space' (Rheingold, 1993). An example of this includes a recent thesis by Atkins exploring the indexing of space through the use of language (mainly deixis) in internet health groups (2011) using a 45,000 word corpus. Before CANELC, a corpus including threads from a wide range of discussion boards, covering a broad spectrum of different topics had yet to be compiled.

4.4. Emails

Emails are often only addressed to specified recipients or groups of readers, and are not outward facing or designed for the public at large, although the number of potential recipients of an email may actually be infinite. In a similar fashion to IM content, users can respond in a chain-like fashion

to previous messages, with as little or as much text as they choose, whenever and wherever they like, via a PC/laptop or mobile phone.

Research into the language of emails is again longstanding; noteworthy examples include works by Baron, 1998, 2000; Crystal, 2001; Danet, 2002 and Panteli, 2002. As with the other e-language types, email corpora are constantly emerging in the research landscape, and the content of the large-scale Enron corpus, most notably, has already been studied in some detail by researchers in this field. Despite such work, the similarities and differences between emails and other forms of e-language, in terms of structural, functional and pragmatic properties remains an underexplored topic. CANELC gives researchers the impetus for carrying out such lines of research, as well as for building on the foundations of what is already known about the language of email. It should be noted, however, that our data consists largely of email collected from business contexts. It is not especially representative. The request from CUP was, however, for us to collect business data to inform text and course book development in business English. A next stage would be to collect a greater variety of less contextually specialised emails.

4.5. Text messages

The final form of e-language included in CANELC is SMS messages (Short Message Services). While 'text messaging was never originally envisioned as a means of communication between individuals…..it was originally conceived of as having commercial use, or possible as a service for mobile phones to signal the arrival of a voicemail message' (Crystal, 2008: 77), 'texting' has become a very central part of communication in modern life, with 11 million text messages being

sent every hour in the UK (as recorded in January 2010)[9].

SMS messages are again more private forms of communication as they are often directed at individuals and small groups of friends. Texting is immediate and often informal. The language of SMS messages has been explored by numerous researchers in linguistics (see Crystal 1998, Döring 2002; Faulkner and Culwin 2005; Grinter and Eldridge 2003; Tagg 2009 and Thurlow & Brown 2003) although, as Crystal notes, 'we are still learning how to behave when we text' (2008: 28). Issues such as when and how one should appropriately respond to messages, if at all, are widely debated. There is thus scope for examining other characteristics of SMS behaviour that are still somewhat underexplored, and again the provision for doing just this is something CANELC offers.

5. Topics covered

CANELC includes data covering a range of different topics. A list of these is provided in figure 4. Originally, it was intended to use pre-existing schema to describe and encode the different topics of the content, but we were unable to find a generic, widely-used classification system for this purpose. Therefore, these categories were defined following extensive discussions by the group of researchers working on the CANELC project. The team looked at the key content words in the descriptions of sites, such as 'food' and 'recipes' in 'Showing the world the beauty of British food and recipes', noted them down for each individual contributor, then attempted to define broad thematic categories based on the key words defined across the dataset. So, for the example just given, text collected from this tweeter would be broadly categorised under the topic of 'cookery', so would be labelled under category 'C'.

The categorisation of the data was carried out semi-automatically. As a first parse, two trained researchers were employed to look at the data and categorise topics manually. The data was then inputted into the semantic tagger of WMatrix (Rayson, 2003) to see whether thematic groupings of the content could be defined using this automated method. Finally, the results from the three stages were compared and lengthy discussions were carried out between the researchers to select which category appeared to be most relevant to specific extracts of data. In situations where differences of opinions could not be resolved, as discussed above, multiple codes were assigned to the data rather than one.

These 'topics' exist on a continuum from the more 'public' concerns (topics in the 'A' category) such as news, politics and current affairs, to the 'private', such as personal and daily life. The entire CANELC dataset has been categorised according to these categories. While the assignment of the content to these categories was fairly transparent in some cases, others were slightly more 'fuzzy', insofar as they discussed multiple topics across the different categories. In these instances the data was given a range of category codes, thus A/B/C rather than simply 'A'.

***Figure 4**:* Topics covered in the CANELC content.

6. Anonymity

To protect the identity of contributors to the corpus and individuals mentioned within it, all content has been fully anonymised. First names (including Twitter IDs etc.) and easily identifiable nicknames were anonymised as [NameX], with 'X' representing a unique number code which is indexed in our metadata files (though these are unlikely to be distributed). Other anonymised

features/codes include the following:

[Address]

(ContentPrivate)

[Bankdetails]

[BusinessName]

[DocumentRef]

[Email]

[FaxNo]

[IPAddress]

[Link]

[Password]

[PhoneNo]

[Photo/Picture]

[PONumber]

[PortNo]

[Postcode]

[ProductName]

[ServerAddress]

[ServerName]

[Signature]

[SoftwareName]

[Sortcode]

[Tagline]

[Username]

[Website]

Anonymising e-language is a challenging process. This is especially true for the shorter and

fragmented contributions such as SMS messages and Tweets. This is because references to persons/ places in such discourse tend to be highly context bound and thus integral to the meaning of the message, making it potentially detrimental to remove them. For this reason, the same approach used by Tagg when developing the CorText corpus was used here, wherein text referring to 'celebrities, film names, public places, characters from film, TV/ reality shows weren't changed' (Tagg, 2009: 98) but references to persons not in the public eye, along with those other features mentioned above, were.

Given that 'popular' blog, discussion board and twitter sites were included in the corpus, such public figures, for example, featured frequently. An example of this is seen in the following tweet:

Sent - 22:17 on 12/12/2010

Content - @[Name1911] Feel exactly the same. Old school Biffy fan, Essex born and Matt fan but I'm conflicted

'Biffy' in this tweet refers to the band Biffy Clyro, whose song was covered by 'Matt', a singer from Essex who won the TV show X Factor in December 2010. Without this extra information, that is the name and identity of the band/singer, the analyst would be unaware of the referential meaning of this tweet. For this reason, details of this nature, such as public figures, designer labels, celebrities, TV programmes/characters and shop names were not anonymised.

To add clarity and extra meaning to such contextually bound referents, an index of unanonymised content was created while constructing the corpus, detailing the name of the referent and who/what

they are. An excerpt from this 'index of cultural referents' is seen in Figure 5.

Figure 5: An example of the index of unanonymised content in CANELC.

Common Christian names and nicknames were also, in some cases, not anonymised because it was felt the identity of the referent could not be easily traced when using such names. An example of this is seen in the following tweet:

Tweet T.12115 @[Name2856] Thx 4 the RT Andy.

It is unlikely that the identity of the specific person 'Andy' can be determined simply by reading this tweet, so it was felt that it would not be cause for concern to leave such names in the data.

7. Storing and representing the data

When permission was granted data was simply extracted from the site(s) or RSS feeds (for Blogs, Discussion Boards and Tweets) and pasted into an XML corpus database. This database was standardised and formatted in the same way as content from the Cambridge English Corpus so that the data can be directly slotted into this corpus.

Data was stored with the following headers, as far as possible, included:

- Author's name, age, gender, nationality
- Date and time composed

- Intended recipient
- Content
- General theme of content
- Follow up comments/ responses
- 'Other' relevant information

Data from emails was forwarded directly to the researcher who could manually input the content into the XML database. Many modern smart phones are compatible with PC based software which allows users to connect their phones via USBs to computers and simply download the content of their SMS messages along with the time and date they were sent. Alternatively, web enabled phones often automatically back-up these details to web accounts which can be downloaded as a database and sent directly to the researcher for use.

8. Analysis

8.1. Key questions

The final part of this paper reports on some preliminary analyses of CANELC. The aim here is to outline some of the basic similarities and differences between the asynchronous data included in CANELC and 1 million word samples of spoken and written language from the BNC[10]. The following key question is examined as part of this analysis:

What are the most frequent words/phrases used in CANELC (within and the across different modes) in terms of word *function*, *sense* and *meaning* and how do these compare to the spoken/written elements of the BNC?

For the purpose of this analysis, non-standardised spellings featured in the corpus, such as *2* (for *to* or *too*), *wanna* (*want to*) and *u* (*you*) were standardised with the help of the software Vard (Baron and Rayson, 2008), prior to being inputted into WMatrix. The corpus data was grammatically and semantically tagged in WMatrix after the standardisation had taken place. Vard enables users to identify spelling irregularities in a corpus then train the system to automatically replace candidates with standardised versions of the words. These were then counted towards cases of the standard spelling of each form. Given that the orthographic formulation of these features was not the primary concern of the analysis, rather the frequency of use with which forms were used, this process of standardisation was deemed sufficient for the needs of the current paper.

8.2. Function

Figure 6 tabulates the top 50 most frequent words and clusters used in the CANELC corpus (note: 'rel. freq.' refers to relative frequency, the frequency of the given word as a proportion of the entire corpus).

Function words, rather than content words, proliferate here. Of these function words we see that, significantly, the use of personal pronouns is shown to be particularly frequent both in the data, with *you*, *I* and *it* ranking highly, along with the demonstratives *this* and *that*. A keyword analysis of these pronouns, in comparison to spoken and written extracts of the BNC corpus (comprising 1 million words each), indicated that their use more closely mirrors spoken forms of discourse, as personal pronouns are characteristically less often used in written language. *I*, for example, was noted to occur once every 38 words in an analysis of some spoken data from the BNC (Leech, 2000) and only

once in every 200 words in the written data. Rates of 1:43 words for the CANELC data are thus far closer to the spoken BNC analysis (this result was also mirrored by Chafe and Danielewicz, 1987; Biber, 1992; Biber et al., 1999; Carter and McCarthy, 2006 and Atkins, 2011).

Figure 6: The most frequent words and clusters used in the CANELC corpus.

As outlined by Heylighen and Dewaele (2003), the frequent use of personal pronouns, along with adverbs, verbs and interjections are typically more characteristic of more informal styles of communication, while nouns, adjectives, prepositions and articles are more frequent in more formal types of language. Based on this crude definition, to provide an insight into the levels of formality across the different communicative modes in the corpus (albeit crudely), figure 7 shows the relative frequencies of each of these parts of speech across the modes in the corpus, compared to the relative frequencies seen in the entire corpus:

Figure 7: Relative frequencies of syntactic categories across the CANELC corpus.

The numbers in bold indicate that the relative frequency for a specific part of speech is lower than that seen across the entire CANELC corpus, while those in italics are higher than the overall relative frequency for the corpus. Blog and twitter data are shown to use parts of speech that are more characteristic of 'formal' language at a higher rate than other modes across the corpus, while discussion boards, emails and SMS messages are closer to more 'informal' language. The most significant differences in relative frequency are seen in the underuse of nouns in the twitter and blog data, the underuse of verbs in the email data and the overuse of verbs in the twitter data.

A variety of reasons may account for these differences, many of which are likely to be closely tied to 'social factors associated with the situation or context of communication' (Herring, 2002: 11 – also see Hymes, 1974 and Baym, 1995). Content sent/received via the blogs and tweets (particularly those selected to be included in this study) is often publicly rather than personally targeted. This means that the readers are often unknown, so the relationship between the blogger/tweeter and reader is often less close than it is with SMS users. The 'purposes of communication' may also be different to emails and SMS messages which, again, in turn affects what they are communicating about and how they achieve this (i.e. the type of language being used). A closer analysis of these social factors and the individual context of communication will allow more specific conclusions about this to be made. Again, the detailed metadata associated with the CANELC data will allow us to explore this further in future studies; there is, however, limited scope to do this here.

SMS and email are often more personal and intended for a bounded number of recipients. The language used in these situations may still be formal, such as in professional emails for example, but the recipient is often more likely to be a known party or somebody within a close proximity of the senders social or peer group.

8.3. Sense and meaning

Figure 8 lists the top 50 key words and clusters used in the CANELC corpus, compared to spoken and written BNC extracts.

Figure 8: Key words and clusters used in the CANELC corpus, compared with spoken and written extracts from the BNC.

Again, *I* is overused in CANELC in comparison to the written BNC data (with a log-likelihood of +5219.64), but there is no significant difference in usage between the CANELC and the spoken BNC data. *You* (rated fourth here with a log-likelihood of +3242.12) and other personal pronouns were all shown to be key words in comparison to the written element but their use was not as statistically different to the use in the spoken BNC data.

Terms related to the general thematic grouping 'information technology and computing' and 'the media, TV radio and cinema' (as characterised by the semantic tagging functionality in WMatrix) such as *Google*, *site*, *twitter*, *blog*, *social*, *media*, *BBC*, and *socialmedia* are also shown to be more common in CANELC than the BNC counterparts. Similarly, references to communicating in digital environments are also particularly common in CANELC, with *fan*, *posted*, *news* and *learning* all featured in the top 50 key words. These latter terms can be broadly categorised under the thematic groupings of 'IT', 'the Media', 'telecommunication' and 'paper documents and writing' (which also includes terms such as *print*, *register*, *delete* and *list*), themes which feature significantly more frequently in CANELC than the other corpora.

Figure 9 reveals some of the other key differences in the semantic categories (based on keywords and phrases used in the data) that are used at a significantly higher rate in the CANELC data, compared to the spoken and written elements of the BNC (based on the UCREL, University Centre for Computer Corpus Research on Language, Lancaster University, semantic analysis system, as featured in WMatrix - see Wilson and Rayson, 1993):

Figure 9: Comparing semantic categories of the CANELC data versus the spoken and written BNC

data.

Content related to time and place (including the categories: geographical names, time: period, time: future, time: present; simultaneous and location and direction) also feature more frequently in the CANELC corpus compared to the spoken and written data. In figure 7 we saw that clusters such as *last night, next week, next year, this morning* and *at the moment,* in particular, are particularly ubiquitous. This is an interesting finding because although e-language is actually asynchronous, there may only be a short delay between the time that e-language is composed and the time that it is read and responded to.

The use of these temporal deictic markers (as with the use of personal pronouns), suggests forms of communication that allow for an immediate or near-immediate information exchange, a forum for communicating reports of events and incidents in near real-time, as the understanding of the temporal referent is shared. In fact, on closer inspection of some of the twitter, email and SMS data in particular, messages were often sent by users and then responded to within hours, even minutes, closing the gap between production and reception, possibly accounting for the differences seen. However, unlike in face to face, spoken discourse, the actual physical space is rarely shared at the point when the message is sent. The lack of shared physical space may lead to an overcompensation in the use of deictic markers, as a means of establishing and reconfirming a shared 'digital space' between senders and recipients. Such reconfirmation is not a required part of spoken interaction as the social, physical and temporal context is frequently changeable.

Aside from deictic markers, figure 9 also reveals that the use of politeness strategies is also more

frequent in CANELC than the BNC data, with log-likelihood score of +1410.74 (a frequency of 103) compared to the spoken BNC and +1303.77 when compared to the written BNC (a frequency of 130). As seen in figure 10, to *thank* someone appears to be a particularly common word used in e-language, occurring 669 times across the corpus.

Figure 10: Common politeness terms use in CANELC.

This frequent use of politeness terms is seen in all sub-types of the CANELC data, with the language of the emails ranking as having a particularly high number of politeness terms and the blogs with the least (although even for blogs, the number is still significantly higher than what is seen in the BNC). This finding mirrors that seen by Herring who found that 'public CMD [computer mediated discourse] tends to be less polite than private CMD' (Herring 2003: 19), although this obviously depends on the purpose of communication, who the message is intended for (i.e. whether it is aimed at a specific person or group of people) and the nature of the relationship between the sender and sendee. The fact that blogs, twitter exchanges and much of the discussion board content included in CANELC is publicly accessible suggests that the maintenance of face and positive politeness are critical ingredients for maximising the number of people that will follow your online existence. This would help to explain the frequent use of politeness strategies across all the modes of e-language examined here. However, specific conventions for doing this 'successfully' are something that needs to be examined in more detail.

Another interesting feature of e-Language, which is used more extensively than in spoken and

written data, is a closing with kisses *x*, *xx*, *xxx*. The average length is a single *x* unless the recipient is defined as a close friend or partner and there was also evidence of the use of *x* between colleagues and friends of the same and different genders (for both men and women), a device which thus seems to be conventionally accepted for use by all. The use of *x* is seen to be highly frequent in all of the modes of e-language; most commonly featuring in SMS messages and least frequently in blog data. *X* broadly functions as a relationship maintenance device, a method of bringing the sender and recipient of a message closer together, again despite the physical distance. It acts almost as a signing off method, more personal and expressive than a full stop or a signature. Again, a more detailed exploration of the differences in usage of *x* across the different e-language modes and individual users is something that will be explored further in the future. Questioning what precedes or follows a message with an *x*, and questioning the function of a message will also help us to construct a more detailed understanding of this feature. For example, compare the following two SMSs from CANELC:

SMS.224

How did the footie go? U watching that drama on 4? Very sad :-(…. Filmed in notts x

SMS.3964

Its just a copy of wots there, theyre usingthe old bits as a template. All in 8x3. They know all this.

The function of the SMS.3964 is purely transactional (send by a manager to a colleague), a form of information provision while the second example is of a more intimate variety, an information request

but in a more socialising capacity. For SMS.224, the *x* acts to maintain and reinforce the relationship between sender and recipient. This is less critical to the second message.

9. Summary

This paper has introduced the one million word CANELC corpus, the Cambridge and Nottingham e-language Corpus. It details how the corpus was constructed and illustrates how it may be used to help us examine the structure and use of language in digital environments with, as can be seen from the corpus construction, opportunities to examine how e-language varies across different domains, across different levels of formality and with particular attention to the spokenness and high levels of interactivity of some e-communication. While further research into discourse within digital domains needs to be carried out, we believe that CANELC provides us with some of the main foundations for doing this.

This opens the door to a variety of interesting questions about the use of language in digital contexts, questions that, with time, we hope to explore further using CANELC. Among the possibilities are: at a more micro-level, analysis of further seemingly e-specific forms such as politeness phenomena and deixis across the database as well as exploration of key recorded forms of spoken grammar outlined by major grammars such as Biber et al (1999) and Carter and McCarthy (2006) including vagueness markers, ellipsis, modal expressions, fronting, headers and tails At a more macro level possibilities include: fuller comparisons between CANELC data and other e-language corpora; the collection and analysis of Facebook data to explore the nature and the extent of linguistic differences and distinctions between this popular medium and other e-language forms;

extending CANELC to embrace a larger range of email data from a wider variety of contexts of use; examining the extent to which spokenness - not just in e-communication but in writing in general- is before our very eyes both a growing phenomenon and a significant part of systemic contemporary language change.

10. References

Allan, S. 2006. *Online News: Journalism and the Internet*. Maidenhead, UK: Open University Press.

Atkins, S. 2011. *A Cognitive Linguistic Perspective on Social Space in Online Health Communities.* Unpublished PhD Thesis. The University of Nottingham.

Baron, N. 1998. "Writing in the Age of Email: the impact of ideology versus technology". *Visible Language 32* (1), 35-53.

Baron, N. 2000. *Alphabet to Email: How Written English Evolved and Where it's Heading.* London: Routledge.

Baron, A. & Rayson, P. 2008. "VARD 2: A tool for dealing with spelling variation in historical corpora". In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Aston University, Birmingham, UK, 22 May 2008.

Bates, E. 1976. *Language and Context*. New York: Academic Press.

Baym, N. 1995. “The emergence of community in computer-mediated communication”. In Jones, S.G. (Ed.) *Cybersociety: Computer-mediated communication and community*. Thousand Oaks, CA: Sage. pp.138-163.

Beißwenger, M. 2007. *Sprachhandlungskoordination in der Chat-Kommunikation (Linguistik – Impulse & Tendenzen 26)*. Berlin. New York: de Gruyter

Biber, D. 1993. “Representativeness in corpus design”. *Literary and Linguistic Computing 8*(4): 243–57.

Biber, D. 1992. “On the complexity of discourse complexity: A multidimensional analysis”. *Discourse Processes 15:*133-163.

Biber, D. Conrad, S. Leech, G. Svartvik, J. & Finegan, E. 1999. *The Longman Grammar of Spoken and Written English.* Harlow: Longman.

Borau, K., Ullrich, C., Feng, J. & Shen, R. 2009. “Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence”. *ICWL 2009*, LNCS 5686, 78–87.

Boyd, D & Heer, J. 2006. “Profiles as Conversation: Networked Identity Performance on Friendster.” In *Proceedings of the Hawaii International Conference on System Sciences (HICSS-39)*, Persistent Conversation Track. Kauai, HI: IEEE Computer Society. January 4 – 7, 2006.

Brown, G. 1989. “Making sense: The interaction of linguistic expression and contextual

information". *Applied Linguistics 10*(1), 97-108.

Carter, R.A. & McCarthy, M.J. 2006. *Cambridge Grammar of English.* Cambridge: Cambridge University Press.

Chafe, W.L. & Danielewicz, J. 1987. "Properties of spoken and written language". In Horowitz, R. & Samuels, S.J. (Eds.) *Comprehending oral and written language.* New York: Academic Press. pp.83-113.

Collot, M. & Belmore, N. 1996. "Electronic Language: A new variety of English". In Herring, S.C. (Ed.) *Computer Mediated Communication: Linguistic, Social and Cross-cultural Perspectives.* Amsterdam: John Benjamins. pp. 13-28.

Condon, S.L. & Cech, C.G. 2001. "Profiling turns in interaction". In *proceedings of the 34th Annual Conference of the Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society Press.

Crystal, D. 1998. *Language Play.* Cambridge: Cambridge University Press.

Crystal, D. 2001. *Language and the Internet.* Cambridge: Cambridge University Press.

Crystal, D. 2003. The joy of text. *Spotlight magazine:* 16-17.

Crystal, D. 2008. *Txtng: the Gr8 Db8.* Oxford: Oxford University Press.

Danet, B. 2002. *The Language of Email*. European Union Summer School lecture, University of Rome, June 2002, Lecture II.

Döring, N. 2002. "1 Brot, Wurst, 5 Sack Äpfel I.L.D – Kommunikative Funktionen von Kurzmitteilungen (SMS)" [1 bread, sausage, 5 bags of apples I.L.Y.' – Communicative functions of text messages (SMS)]. *Zeitschrift für Medienpsychologie,* 14 (3), 118–128.

Duranti, A. & Goodwin, C. (Eds.). 1992. *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press.

Faulkner, X. & Culwin, F. 2005. "When fingers do the talking: A study of text messaging". *Interacting with Computers 17*(2): 167-185.

Forsyth, E.N. & Martell, C.H. 2007. "Lexical and Discourse Analysis of Online Chat Dialog". In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, September 2007, 19-26.

Gillmor, D. 2004. *We the Media: Grassroots Journalism By the People, For the People.* Farnham, Surrey: O'Reilly Media.

Green, L. J. 2002. *African American English: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Grinter, R. E. & Eldridge, M. 2003. "'Wan2talk? Everyday text messaging". *Proceedings of the*

ACM Conference on Human Factors in Computing Systems (CHI): 441-448.

Halliday, M.A.K. & Hasan, R. 1989. *Language, Context and Text: Aspects of language in a Social Semiotic Perspective.* Oxford: OUP.

Hard af Segerstag, Y. 2002. *Use and Adaptation of the Written Language to the Conditions of Computer-Mediated Communication*. Unpublished PhD thesis. University of Goteborg.

Herring, S.C. 2002. "Computer-mediated communication on the Internet". *Annual Review of Information Science and Technology 36*: 109-168.

Herring, S.C. 2007. "A faceted classification scheme for computer-mediated discourse". *Language@Internet 4*(1): 1-37.

Heylighen, F. & Dewaele, J. -M. 2003. "Variation in the contextuality of language: an empirical measure". *Foundations of Science* 7: 293–340.

Honeycutt, C., & Herring, S. C. 2009. "Beyond microblogging: Conversation and collaboration via Twitter". *Proceedings of the Forty-Second Hawaii International Conference on System Sciences (HICSS-42)*. Los Alamitos, CA: IEEE Press.

Horn, C., Pimas, O., Granitzer, M., Lex, E. & Graz, K-C. 2011. "Realtime Ad Hoc Search in Twitter: Know-Center at TREC Microblog Track 2011". In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*, Gaithersburg, Maryland, November 15-18, 2011.

Hymes, D. 1974. *Foundations in sociolinguistics: An ethnographic approach.* Philadelphia: University of Pennsylvania Press.

Jansen, B.J., Zhang, M., Sobel, K. & Chowdury, A. 2009. "Twitter power: Tweets as electronic word of mouth". *Journal of the American Society for Information Science and Technology 60*(11): 2169 – 2188.

Jones, Q. 1997. "Virtual-communities, virtual settlements and cyber archaeology: A theoretical outline". *Journal of Computer Mediated Communication 329*, 3.

Klimt, B. & Yang, Y. 2004. "Introducing the Enron Corpus". In *Proceedings of CEAS 2004 - First Conference on Email and Anti-Spam*, Mountain View, California, USA, 30-31.

Ko, K. 1996. "Structural characteristics of computer-mediated language: A comparative analysis of InterChange discourse". *Electronic Journal of Communication 6*(3).

Labov, W. 1972. *Sociolinguistic Patterns.* Philadelphia, PA: University of Pennsylvania Press.

Leech, G. 2000. "Grammar of spoken English: new outcomes of corpus-oriented research". *Language learning 50*(4): 675-724.

Ling, R. 2003. "The socio-linguistic of SMS: An analysis of SMS use by random sample of Norwegians". In Ling, R. and Pedersen, P. (Eds.) *Mobile communications: Renegotiation of the social sphere.* London: Springer. pp.335-349.

Myers, B. 2010. “Theology 2.0: Blogging as Theological Discourse”. *Cultural Encounters* 6 (1): 47–60.

Myers, G. 2010. *The discourse of blogs and wikis*. London: Continuum.

Nelson, K., Engel, S. & Kyratzis, A. 1985. “The evolution of meaning in context”. *Journal of Pragmatics 9*: 453-474.

Oksman, V. & Turtianen, J. 2004. “Mobile communication as a social stage: meanings of mobile communication in everyday life among teenagers in Finland”. *New Media and Society* 6 (3): 319-339.

Orasan, C. & Krishnamurthy, R. 2002. “A corpus-based investigation of junk emails”. In *Proceedings of LREC-2002*, Las Palmas, Spain.

Panteli, N. 2002. “Richness, Power Cues and Email Text”. *Information and Management 40:* 75–86.

Puschmann, C. 2009. “Diary or Megaphone? The pragmatic mode of weblogs”. Paper presented at *Language in the (New) Media: Technologies and Ideologies*, September 3-6 2009, Seattle, WA, USA.

Rayson, P. 2003. *Matrix: A Statistical Method and Software Tool for Linguistic Analysis Through Corpus Comparison* . Unpublished PhD thesis. Lancaster University.

Rheingold, H. 1993. *The Virtual Community: Homesteading on the Electronic Frontier*. New York: HarperCollins.

Schler, J., Koppel, M., Argamon, S. & Pennebaker, J. 2006. "Effects of Age and Gender on Blogging". In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Scollon, R. & Scollon, S. 2003. *Discourse in Place: Language in the Material World*. London: Routledge.

Shortis, T. 2007. "Gr8 Txtpectations: the creativity of text spelling". *English Drama Media Journal* 8, 21-26.

Sutherland, J. 2002. "Cn u txt?". Featured in *The Guardian*, 11th November.

Tagg, C. 2009. *A Corpus Linguistics Study of SMS Text Messaging*. Unpublished PhD Thesis. The University of Birmingham.

Thurlow, C. & Brown, A. 2003. "Generation Txt? Exposing the sociolinguistics of young people's text-messaging". *Discourse Analysis Online 1*(1).

Widdowson, H.G. 1998. "Communication and Community: The Pragmatics of ESP". *English for specific purposes 17*(1): 3–14.

Wilson, A. & Rayson, P. 1993. "Automatic Content Analysis of Spoken Discourse". In Souter, C. & Atwell, E. (Eds.) *Corpus Based Computational Linguistics*. Amsterdam: Rodopi. pp.215-226.

Zappavingna, M. 2011. *Discourse of Twitter and Social Media How We Use Language to Create Affiliation on the Web.* London: Continuum.

[1] School of Education, Communication and Language Sciences, Newcastle University, Newcastle, NE1 7RU

Correspondence to: Dawn Knight, *e-mail*: Dawn.Knight@ncl.ac.uk

[2] School of English Studies, University Park, The University of Nottingham, Nottingham, NG7 2RD

[3] CANELC stands for Cambridge and Nottingham e-language Corpus. This corpus has been built as part of a collaborative project between The University of Nottingham and Cambridge University Press with whom sole copyright of the annotated corpus resides. CANELC comprises one-million words of digital English taken from SMS messages, blogs, tweets, discussion board content and private/business emails. Plans to extend the corpus are under discussion. The legal dimension to corpus 'ownership' of some forms of unannotated data is a complex one and is under constant review. At the present time the annotated corpus is only available to authors and researchers working for CUP and is not more generally available.

[4] CANCODE stands for *Cambridge and Nottingham Corpus of Discourse in English*. This corpus has been built as part of a collaborative project between The University of Nottingham and Cambridge University Press with whom sole copyright resides. CANCODE is comprised of five-million words of (mainly casual) conversation recorded in different contexts across the British Isles.

[5] CEC stands for Cambridge English Corpus, a corpus of over one billion written and spoken words in English. For more information visit: http://www.cambridge.org/

[6] Externally commissioned research is to some degree always subject to the requirements of the agency that commissions the research and the balance of CANELC data is determined accordingly with SMS and email datatypes assuming a smaller proportion. The next phases of the research may indeed see each of the data-type categories balanced more evenly. However, SMS and email data are categorised by a markedly interpersonal dimension and when aggregated do constitute a further balancing category in the whole corpus.

[7] Examples of such sites include the following: www.guardian.co.uk/technology/2008/mar/09/blogs, http://modernl.com/article/uk-blogosphere-top-10-british-blogs, http://wefollow.com/twitter/british, www.Britishblogs.co.uk, www.telegraph.co.uk/technology/twitter/6832287/Most-influential-British-twitter-users-revealed.html,

[8] Facebook status updates are not included CANELC as the fact they can be viewed and commented on by an array of different users ('friends'), commonly commenting on private information about the user and his/her friendship group, brought into question concerns about copyright and data ownership. To avoid problems with access, ethics and copyright, this data was not included.

[9] Figure taken from findings of The Mobile Data Association: http://www.themda.org/ (accessed 1-6-11).

[10] The British National Corpus, BNC, is a 100 million word corpus of written and spoken discourse in English. For more information see: http://www.natcorp.ox.ac.uk/