e-Language: Communication in the Digital Age

Dawn Knight[1], Newcastle University

1. Introduction

Digital communication in the age of 'web 2.0' (that is the second generation of in the internet: an internet focused driven by user-generated content and the growth of social media) is becoming ever-increasingly embedded into our daily lives. It is impacting on the ways in which we work, socialise, communicate and live. Defining, characterising and understanding the ways in which discourse is used to scaffold our existence in this digital world is, therefore, emerged as an area of research that is a priority for applied linguists (amongst others).

Corpus linguists are ideally situated to contribute to this work as they have the appropriate expertise to construct, analyse and characterise patterns of language use in large-scale bodies of such digital discourse (labelled 'e-language' here - also known as Computer Mediated Communication, CMC: see Walther, 1996; Herring, 1999 and Thurlow et al., 2004, and 'netspeak', Crystal, 2003: 17).

Typically, forms of e-language are technically asynchronous insofar as each of them is 'stored at the addressee's site until they can be 'read' by the recipient (Herring 2007: 13). They do not require recipients to be present/ready to 'receive' the message at the same time that it is sent, as spoken discourse typically does (see Condon and Cech, 1996; Ko, 1996 and Herring, 2007). However, with the increasing ubiquity of digital communication in daily life, the delivery and reception of digital messages is arguably becoming increasingly synchronous.

Mobile apps, such as Facebook, WhatsApp and I-Message (the Apple messaging system), for example, have provisions for allowing users to see when messages are being written, as well as when they are received and read by the. This is helping to create a shared digital, rather than physical, space between the sender and recipient of the message, making a near-immediate exchange of communication (see Knight et al., 2014). This creates, a blurring of the boundaries between what we traditionally understand as being characteristic of spoken and written discourse through the reduction of the temporal and social distance of between the sender and receiver. This acknowledgement of this 'blurring' has provided the impetus for the focus of the current chapter, with the examination of the use of a particular linguistic feature, modal verbs, in e-language in comparison to speech and writing.

To date, the majority of work in corpus linguistics on the description of e-language has focused on using either small-scale or bespoke corpora, so corpora designed to either meet and answer a specific linguistic enquiry, or corpora containing only one data-type and/or e-language variety (see Klimt and Yang, 2004; Schler *et al*., 2006; Beißwenger, 2007 and Tagg, 2009). While there are many advantages associated with using specialist and/or small-scale corpora, there are few corpora in existence which allow users to comment on e-language use in general. This has meant that the ways in which we live and communicate in the digital world 'across multiple resources, remains an under-explored area of research in corpus linguistics' (Knight et al., 2013: 30).

The CANELC corpus[2] (Cambridge and Nottingham e-Language Corpus, see Knight et al., 2013) attempts to fill this 'gap'. It allows users to query data at a general level, as well as across the 'genre' (Herring, 2002) of e-based communication used (e.g. blogs, emails, tweets

[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] 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 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.

etc.). While such genres may comprise different 'socio-technical' modes, so are likely to have 'social and cultural practices that have arisen around their use' (Herring, 2007: 3), CANELC was constructed on the premise that there are likely to be some key similarities between each of them on the basis of them all being e-based communication systems. This justifies their inclusion in the same 'general' corpus (much the same with 'general' written and spoken corpora, which include language from different text-types, genres and contexts of communication).

This chapter examines how we can start to build better descriptions of e-based discourse through the analysis of real-life examples of mixed source e-language, as evidence by corpora. Discourse is defined here as language-in-use in digital contexts, observed from both a micro (i.e. word-by-word, sentence and text-by-text level) and macro (i.e. 'beyond the text', considering the more socio-ideological factors influencing language choice and use) perspective. This chapter focuses specifically on exploring the incidence and frequency of modal verb usage in CANELC, and compares this to written and spoken samples of language taken from the BNC[3] (British National Corpus). Based on these analyses, questions as to whether e-language appears more or less (in)direct and/or (im)polite than spoken and written discourse are explored.

2. Modality

2.1. Functions and forms

A marker of modality is used to 'refer to a speaker's or writer's attitude towards, or point of view about, a state of the world' (Carter and McCarthy, 2006: 638). As described by Palmer (1979) modality is typically expressed through the use of modal verbs and associated forms that are utilised to make subjective judgements about the truth, certainty or probability of a proposition (*epistemic*); whether something is speculative or more definitive (*dynamic*) and the ability, obligation or duty for carrying out the proposition expressed in an utterance (*deontic*).

Often specific lexical forms of modal verbs are multifunctional, so have the potential to express more than one kind of modality, depending on the context/co-text in which they are used. For example:

(1) Sometimes its nice to be observer not entertainer. Dont worry, wont say a word. We ***must*** have a chopsy soon x

(2) It just dawned on me, I ***must*** be the only geek tht didn't ask for any tech for xmas…..

In (1) (taken from the SMS sub-corpus of CANELC), *we must* is used as an order or to provide a sense of obligation: *we must have a chopsy* (chat/ catch-up) *soon*, while *must* in (2) (taken from the twitter sub-corpus) is instead a conjecture as it presents a proposition, *I must be the only geek that didn't ask for any tech,* is a statement which is unproven and impossible to ever fully test/prove (i.e. this use of the modal verb relates to the truth/probability of the proposition).

Modality markers have a highly interpersonal function and are used to mark personal relationships. They also often function as hedging devices that are 'expression[s] of tentativeness and possibility' (Hyland, 1996: 433) and convey politeness, indirectness and assertiveness in discourse, operating to 'mitigate the directness of what we say' (O'Keeffe et al., 2007: 174 – for more information on politeness theory and the notion of 'face', see Brown and Levinson 1978). An example of this follows (taken from the SMS sub-corpus in CANELC):

[3] British National Corpus, BNC, is a 100 million word corpus of written and spoken discourse in English (90% written, 10% spoken). For more information see: http://www.natcorp.ox.ac.uk/

(3) Let me know what you fancy. If our tastes differ then I guess we ***could*** go our separate ways some of the time. I'm interested to know which ones you're interested in though.

I guess we could functions as a hedge here. The speaker is making a suggestion for action, but through indirect means, stressing that this is an option that the receiver may wish to take up, rather than using a directive, saying that they specifically 'should' go their own separate ways. Here the speaker doesn't want to impinge on the receiver so uses this tactic to save face. Interestingly, in this example, the follow-up statement reveals that the speaker would perhaps prefer that the receiver agrees that they *could* go together rather than their separate ways, as the speaker is *interested* in knowing what the receiver would like to see/do.

As well as being multifunctional, Carretero notes that 'modal expressions are [also] grammatically [and lexically] diverse' (1992: 18) as while specific 'core' modal verbs exist (see Table 1), modal expressions are not fixed to these forms. They can also comprise parenthetical expressions (*I think*), adjectives (*probably*), adverbs (*perhaps*), indefinite adjectives or pronouns (*something*), tag questions (*it was Tim, wasn't it?*), hedges (*kind of*) and even contradictions. Perhaps for this reason, the methods and approaches for classifying the forms and functions of modality markers are numerous in linguistic theory (for examples see Palmer, 1979; Halliday, 1985; Biber et al., 1999; Portner, 2009).

For ease of reference, the general list of common modal forms provided by Carter and McCarthy (2006: 427-429) will be used as the basic reference point for the present chapter, with the analysis focusing on the 'core modal verbs' alone:

Type	**Examples**
Core modal verbs	*can, could, may, might, will, shall, would, should, must*
Semi-modals verbs	*dare, need, ought to, used to*
Verbs which can express modal meaning	*hope, manage, suppose, seem, wish, want*
Modal phrases that have become lexicalised	*had better, be meant to, be obliged to, be supposed to*

Table 1: Common modal forms in English (based on the CEC – Cambridge English Corpus[4]).

As e-language is perhaps a hybrid of spoken and written communication, it would be naïve to assume that standard orthographic forms of modal verbs (as are common in spoken and written corpora) are all that exist in this form of communication. Consequently, special considerations must be made when examining e-language. To account for this, the incidence of non-standard forms of modality markers were noted (and tagged) as the CANELC data was manually anonymised. Manual anonymisation is a lengthy process, but was employed here to ensure that ethical prescriptions for this data were met, and to enable the data to be integrated into CEC. Tagging these items meant that they could be located in the same way as standard spellings, when specific terms were searched for.

Alternative methods for carrying out this process include the use of VARD (Baron and Rayson, 2008). , software which enables users to identify spelling irregularities in a corpus then train the system to replace candidates with standardised versions of the words automatically (to enable statistical analyses to be carried out)In this chapter, non-standard spellings of the core modal verbs including *cud*, *wud*, *mite* and *shud* were standardised and are included in

[4] This list is based on evidence from the Cambridge English Corpus (CEC). The CEC contains over one-billion written and spoken words in English: http://www.cambridge.org/

frequency counts for *could*, *would*, *might* and *should* in the analysis . Given that the specific orthographic formulation of these features is not the primary concern of this paper this approach was deemed to be a legitimate one to use.

2.2. Corpus-based studies of core modal verbs

A wealth of corpus-based research has been carried out into the diachronic use of modal verbs, that is, mapping patterns of their frequency of use over time (see Coates, 1995; Biber et al., 1999; Krug, 2000; Nuyts, 2006 and Bowie et al., 2013). Leech et al.'s, (2003 – also Leech et al., 2009), study of modal verbs in 4 spoken and written American (LOB - 1961 and FLOB - 1991) and British English corpora (Brown, 1961 and Frown, 1991), for example, revealed a 12.2% and 9.5% decline in use across these varieties of English over time. This pattern was evident for the majority of individual modal forms (*would, will, can, could, may, should, must, might, shall, ought (to)* and *need(n't)*), with the only outliers to this decline being the use of *can* and *could* in British English (although the difference in use was only +2.2% and +2.4% for these).

A study by Millar, however, provided evidence that stood in opposition to this, claiming instead that while some forms have fallen in frequency (*must*, *shall* and *ought to*), there is a general growth rather than decline in modal use. In contrast to Leech et al.'s use of two data points (a 'snapshot' of data – Millar's main criticism of this work) in providing the basis for reporting a 'general trend', this study focused on data from the TIME magazine corpus, which contains over 100 million words from TIME magazine, from each issue between 1923 and the present day. This corpus arguably allowed for a finer-grained year-on-year analysis of change. Interesting, while a comparison of a small 30 year snapshot of this data (1962 vs. 1991), yielded similar results to Leech, a more detailed analysis over the entire time period suggested that there was in fact a 22.9% increase in the use of core modal verbs over the entire time period.

In response to Millar's criticisms, Leech expanded his study in 2011, by examining a wider range of data (written American and British English from 1901+). Preliminary analyses of this data still indicated that a decline of individual modals such as *may*, *shall* and *must* did occur over the entire time period, contrary to Millar's findings. While such results did provide some strong evidence to support the notion that there is a decline in use across language in its entirety, this 'does not mean that modals are all declining in use at the same rate, or even that all core modals are declining in use' (Bowie et al., 2013: 79) as the rate of use fluctuates significantly from one lexical form to the next. So while an increase was seen in a specific variety/genre of language represented by the TIME magazine text, this is perhaps not indicative of language as a whole where, instead, a general decline has been seen.

Research into the synchronic use of modal verbs, that is, across text type, speakers (e.g. L1 vs. L2 speakers of English – see Hinkel, 1995), varieties of English (Krug, 2000; Smith, 2003; Berglund, 1999 and Bowie et al., 2013), genre and context (including spoken vs. written discourse - see Kiefer, 1987; Benincà and Poletto, 1997; Aijmer, 2002; Narrog, 2005 and Hansan and de Haan, 2009) also indicates that the use of specific modality markers in discourse is highly variable and perhaps text/genre and context bound.

Bowie et al.'s study, for example, noted that the modal verbs *can, could, would* were found to be more frequent in spoken English whilst *may, must* and *should* were less frequent (2013). A similar pattern to this was found in this study when comparing 'non-printed' (i.e. less formal) texts with printed texts.

In a similar vein, O'Keeffe et al. (2007) carried out a study which sought to compare modal usage from fiction texts, taken from a sub-corpus of the BNC; newspaper texts from the CIC (Cambridge International Corpus – now known as the CEC); and a 12 million word academic sub-corpus, also taken from the CIC. Results revealed that a there was a dramatic difference in the types of modals used across these text-types, with *would* and *could* featuring

most frequently in the fiction, *will* in newspapers and *may* in academic texts (see Hewings and Hewings, 2004 and Hinkel, 2009 for additional studies of modality in academic writing). These findings mirrored those gained by Biber et al.'s, 1999 study (also see Hinkel, 2009) of the Longman Grammar of Written and Spoken English (LGWSE), which found that while in general, modal verbs are most common in conversation and least common in news and academic prose, the use of *may* with the meaning of possibility is more common in academic writing than conversation.

In general then, while the use of modal verbs is somewhat dynamic and changeable, their use is highly dependent on the discursive context and co-text in which they are used. The use of certain forms of modals is seen to be particularly characteristic of spoken, informal discourse, fiction and interpersonal encounters (including the forms *could* and *would*) in which the use of these devices help to 'downtone…..the force of an utterance for various reasons e.g. politeness, indirectness, vagueness and understatement' (Farr et al., 2004: 13). In more formal, transactional encounters (e.g. shop encounters – see Farr et al., 2004, or news reports and academic prose) the use of modal verbs is reportedly less frequent as the need to protect face and mitigate communication in such environments less prominent than in more informal, interpersonal environments (e.g. general spoken discourse). The current chapter adds to this literature by examining the use of modal verbs in e-language, providing an insight into how their use in these emergent forms of communication compare/contrast to spoken and written discourse.

The levels of formality in particular forms of e-language has already received attention from a range of researchers (e.g. Crystal, 2008; Shortis, 2007 and Sutherland, 2002) who have identified that, for example, emails and SMS messages are more informal and 'speech-like' (Tagg, 2009: 17) than traditional forms of written language, while the frequent use of certain forms of 'content words' such as nouns, adjectives, prepositions and articles in blogs and tweets (Knight et al., 2013: 47) suggested that these forms were more aligned with more formal, written discourse. This chapter works on the premise that there will be similar variation in the use of markers of modality across the different forms of e-language in CANELC. There is an expectation that their use across all forms of e-language (i.e. the corpus in its entirety) will perhaps be more frequent than in written discourse, but less frequent than spoken, although perhaps blogs, for example, will show an decreased use in these terms when compared to SMS messages and discussion board threads, where the rate of use will be more closely aligned with more interpersonal (but informal), spoken discourse.

3. CANELC

The chapter provides a corpus-based analysis of core modal verb forms used in 972,773 words taken from the one-million-word CANELC corpus (25% from blogs, discussion boards and emails and 10-15% from SMS (Short Message Service) and tweets[5]). The data sample used in this study includes 253,313 words from blogs, 123,291 from emails, 232,759 from discussion board threads, 264,496 from tweets and 98,913 words from SMS messages.

In addition to message content, CANELC contains detailed metadata about the age, location, occupation of the sender and receiver of a message (where available). This provides users with the means for, at least partially, reconstructing elements of the context in which the language was originally used, to provide a basis for explaining why particular patterns of use may exist. The corpus is also classified in terms of the genre/topic(s) covered by each contribution, as seen in Table 2.

[5] Externally commissioned research is to some degree subject to the requirements of the agency that commissions the research and the balance of CANELC data is determined accordingly with SMS and email data types 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.

These topics, very crudely, exist on a continuum with more 'public' concerns such as news, politics and current affairs at one end and more 'private' matters, such as personal and daily life, at the other. This system of classification provides an additional point of entry for analysing CANELC, allowing users to explore patterns of language used when discussing particular topics or genres, within and across the different sub-corpora (providing accurate descriptions of what is seen).

Public ⟷ **Private**

A	B	C	D	E	F
News, media & current affairs	Culture, literature & the arts	Technology, computing & gaming	Music	Celebrity news & gossip	Health & beauty
Weather & the environment	Fashion	Hobbies & past times	Sport	TV	Parenting & family life
Business & finance	Teaching & education	Travel		Humour	Personal & daily life
Politics		Cookery			

Table 2: Topics covered in CANELC.

This continuum mirrors Crystals' conceptualisation of spoken and written discourse as existing on a continuum of formality, with the more formal structures and conventions positioned at the public/written end and the least formal towards the private/spoken end (2003: 17). Knight et al.'s 2014 study started to examine where e-language is positioned on this continuum, and the current chapter continues to explore this.

4. Methodology

This chapter focuses on exploring the frequency of core modal verb usage in CANELC:

- Across different data-types (i.e. sub-corpora).
- Across different text-types (i.e. genres/topics).
- Within/across the different data-types in comparison to spoken and written elements of the BNC.

It utilises a data-driven corpus-based approach as a means of querying data in CANELC. The typical 'way-in' to the analysis of corpora is through generating frequency lists, to determine how frequent particular word forms are across either an entire corpus or across particular sub-corpora (i.e. data-types or text-types in this instance), to allow comparisons to take place. This is complemented by the use of relative frequencies in this chapter, denoting the number of times a specific search term (i.e. 'word') is used at a 'per word' rate in a given (sub)corpus. Log-likelihood (LL) scores are also used to provide a basic statistical measure of the relationship between frequencies, indicating whether specific patterns of similarities or differences are likely to exist by chance or not (with a p value of <0.01 - so a critical value would be in the range of >6.63). Thus, in the next section, a '+' log-likelihood score is used to indicate that a particular rate of use is statistically higher in the first cited variable compared to the other variable under study, and a '-' score is the reverse of this. Rayson's WMatrix software tool (2003) is used to help carry out these enquiries.

The reference corpus with which this data is compared is the BNC, which despite being built in the 1990s, is still considered to be one of the most representative corpora of the English language. It is also balanced, containing written and spoken samples of language from a range

of different text types and discursive contexts, so is a useful point of comparison to CANELC. The data utilised in this study comprises the entire BNC, which comprises 86,299,736 words of written data and 9,963,663 words of spoken data.

It is important to note that to carry out the analysis in this chapter, all cases of modal forms that were in fact proper nouns were not included in the results. Negative forms of words were not included in these analysis only the most frequent positive forms detailed above.

5. Analysis

5.1. General patterns across all modal verb forms

Table 3 charts the raw and relative frequencies (i.e. the number of times the search term (i.e. 'word') is used at a 'per word' rate in the entire corpus) of modal verb usage across each of the sub-corpora (data-types) in CANELC and the spoken and written BNC, while Figure 1 compares these lexical items, charting the LL comparisons of their frequencies of use across all data-types (and CANELC as a whole):

	Spoken	**Written**		**CANELC**		**Blogs**	**Discussion boards**	**Emails**	**SMS**	**Tweets**
Raw freq.	143746	1089530		14904		3303	3723	3369	1772	2664
Relative freq.	1.44	1.26		1.53		1.3	1.6	2.73	1.79	1

Table 3: The frequency of core modal verb usage in CANELC and the BNC.

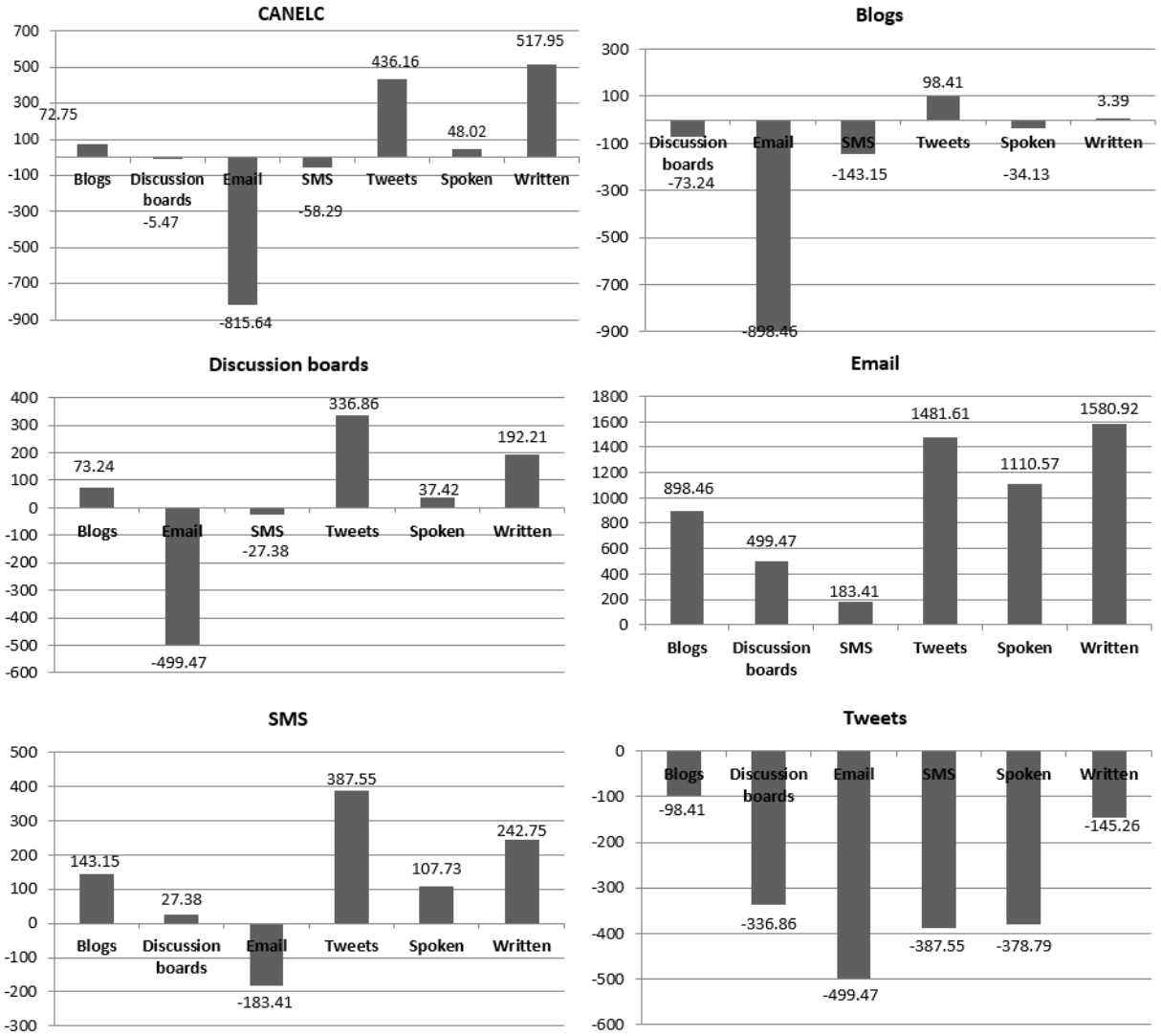

Figure 1 Log-likelihood comparisons of core modal verb forms across the different data-types in CANELC and the BNC.

As with previous studies, Figure 1 and Table 3 support the finding that the rate of modal verb usage in spoken discourse is higher than in written, although the frequency of use of individual forms is variable (discussed below). We also see that the rate of use across CANELC (as a whole) is slightly greater than the use in the spoken BNC (with relative frequencies of 1.53 and 1.44 respectively, with a LL of +48.02), with the rate of use proving to be significantly more frequent in the emails, SMS and discussion board content than in both the spoken and written components of the BNC.

The twitter data sees a rate of modal verb use that is statistically lower than the spoken and written BNC (LL of -378.79 and -145.26 respectively), as well as all other data-types included in CANELC. This is an interesting finding as twitter is near-synchronous and highly interpersonal, often with tweets directed at specific individuals or groups of individuals with a shared interest or opinion. It therefore would perhaps be expected that there would be a necessity to save face and attempt to retain these (virtual) relationships in tweets, but the infrequent use of modality markers suggests this is not the case. A reason for this is perhaps that tweets are restricted to 140 characters, so the need for the economy of expression

potentially strips the language of features which act as hedging devices, mitigating the directness of what is being said.

The most significant difference in usage exists between the tweets and emails, with a LL score of -1481.61 for the tweets. This is an interesting finding, although one which is not overly surprising given that many of the emails included in CANELC were actually gathered from professional, business contexts. There are specific recipients to whom the messages are directed, but the nature of the relationship between the sender and recipient is a formal, working one. It would be interesting to extend this line of enquiry by comparing a corpus of purely social emails, as it is expected that a likely decrease in the number of modal markers will be seen.

Similarly, the rate of use in the written BNC and blogs is generally statistically lower than in all other data-types (aside from the tweets), and have the closest relationship in terms of overall modal verb usage than other text types (with a LL of only -3.39 difference of use in the written BNC compared to the blogs, which is not statistically significant).

Conversely, we see that the use of core modals in the email data occurs at significantly higher rate than in the other CANELC sub-corpora (with a relative frequency of 2.73, which is also much higher than CANELC as a whole, which stands at 1.53) and in comparison to the spoken and written BNC data. As seen in Figure 1, LL scores of this difference range between +183.41 and +1580.92. The use of modals in SMS messages is also statistically greater than the other text types (aside from the emails), with the smallest LL difference seen against the discussion boards (LL +27.38), and the biggest difference when compared with the twitter data (LL +387.55). In an attempt to explain these findings, it is useful to now consider patterns of use for the modals separately.

5.2. Modal verbs forms across text types

Figure 2 charts the relative frequencies of the individual forms of core modal verbs used in CANELC and the BNC sub-corpora, while Figure 3 presents the relative frequency of use across the different data-types in CANELC.

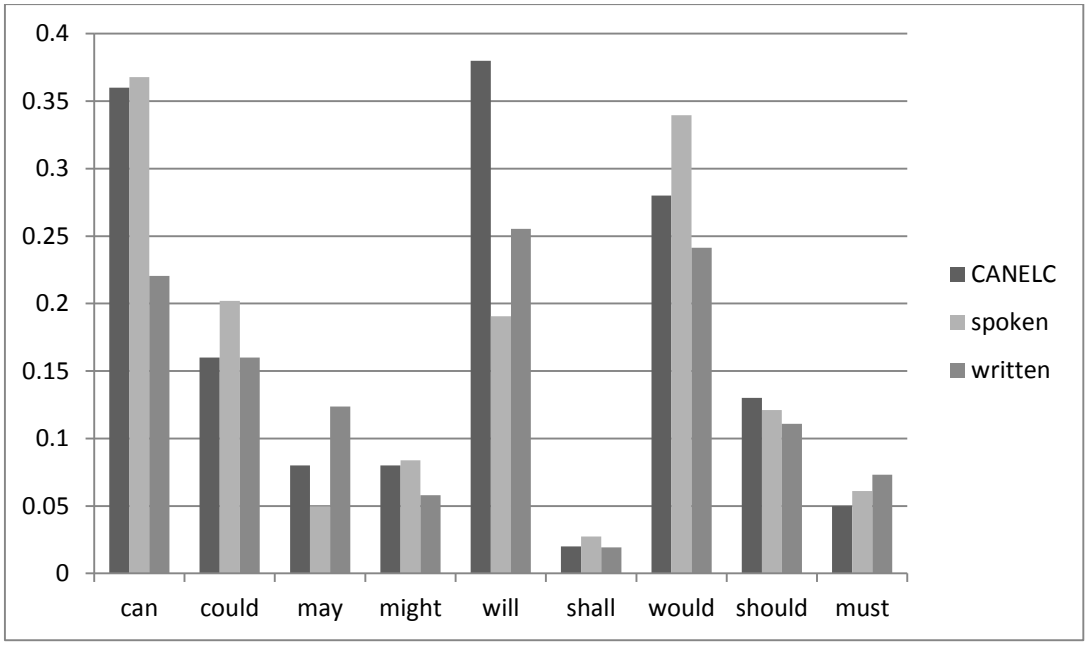

Figure 2: Relative frequencies of core modal verb use in the spoken and written BNC.

Figure 2 reveals that *can*, *will*, *would* and *could* are the most frequently used modal verb forms overall (although *could* ranks higher than *would* in the spoken BNC). *Shall*, *must*, *may* and *might* are least frequently used, with *shall* being used with a relative frequency of ≤0.02 across all (sub)corpora analysed (at a raw frequency of only 197 in CANELC and 16672 out of all 97236172 words examined in the BNC).

The modal verbs *will*, *may* and *should* are more frequently used in CANELC than in the spoken BNC (with LL scores of +1262.22, +108.61 and +13.64 respectively), with *would, could*, *must* and *shall* occurring at a higher rate in the spoken BNC than CANELC (at LL +103.63, +81.6, +56.17 and +18.57 respectively). There is no statistical difference in the use of the forms *might* and *can* across the spoken BNC and CANELC, nor *shall* and *could* in the written BNC and CANELC. Conversely, the forms *can*, *would*, *will*, *might*, *would* and *should* are statistically more frequent in CANELC than the written BNC, while *may* and *must* are statistically more frequent in this latter dataset than the former (with LL scores of >-/+35). *Can*, *would*, *could*, *might*, *shall* and *should* are all statistically more frequently used in the spoken BNC than the written BNC, while *may*, *will* and *must* are more frequent in the latter.

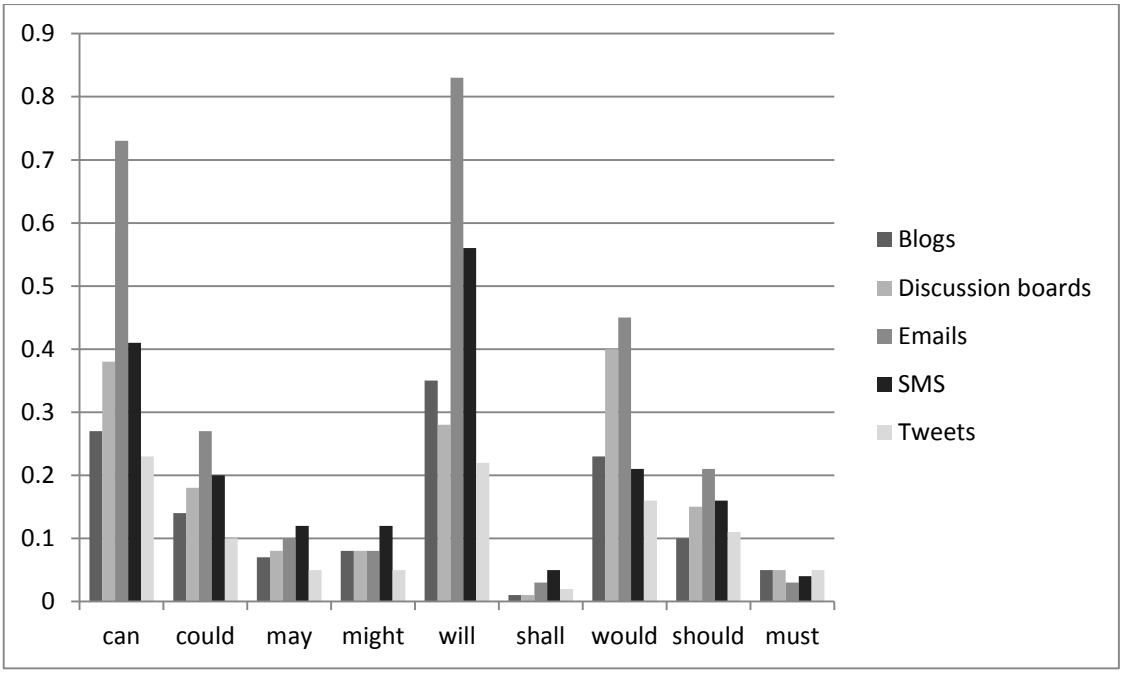

Figure 3: Relative frequencies of core modal verb use in CANELC.

As seen in Figure 3, *can* appears as the second most frequent core modal verb across all data-types in CANELC (this is also true for the spoken BNC, while it is the 3rd most frequently used form in the written BNC) aside from tweets, where it is the most frequent form (with a relative frequency of 0.23). *Will* appears as the most frequent form used in blogs, SMS and emails (and most frequent across the entire corpus and in the written BNC), while it is ranked 3rd in discussion boards and 2nd in tweets. *Would* is the most frequently used form in the discussion board data (as with the spoken BNC – with relative frequencies of 0.4 and 0.37 respectively), but is only 3rd most frequent in the other sub-corpora. *Shall* and *must* appear in the bottom two in terms of frequency of use, across all text types (each with a relative frequency of ≥0.05) aside from tweets where *must* is 3rd from bottom (i.e. 7th most frequent) and *may* is 2nd from bottom.

The use of *must* proved to be statistically more frequent in the written BNC than in any other data type. While *may* was used at a more frequent rate in the written BNC than the spoken BNC, CANELC data-types aside from SMS messages (where no statistical difference was seen

– LL +0.04) and the blog data where it was used at dramatically lower rate (LL -540.24). In terms of the statistical significance of these patterns of use, we see that email data, and to a certain extent, the SMS data, have a general tendency to use the core modal verb forms at a more frequent rate than in the other text types. This is shown in Table 4 (those in bold are used at a less frequent rate in these sub-corpora, but at a rate lower than >6.63, so the difference is not significant):

	Emails					**SMS**			
	can	***could***	***will***	***should***	***would***	***might***	***may***	***shall***	***will***
Blogs	383.9	69.1	357.47	75.7	118.3	11.16	21.93	38.74	75.15
Discussion Boards	195.95	28.22	493.8	17.06	**4.82**	15.3	9.84	49.78	141.79
Emails	N/A	N/A	N/A	N/A	N/A	11.48	**3.12**	**2.8**	-58
SMS	102.65	10.3	58	14.33	94.07	N/A	N/A	N/A	N/A
Tweets	509.32	133.88	689.56	55.05	255.67	40.46	44.83	12.19	237

Table 4: LL comparisons modal verbs in the email and SMS data compared to the other data-types in CANELC.

Conversely, in Table 5 we see that the tweets have a general tendency to use the core modal verb forms at a less frequent rate than in the other text-types. The exceptions this are the modal *must* which is used at a statistically similar rate across all of the text-types aside from the twitter vs. email data where there is a LL score of -9.13; *shall*, where the LL score is +12.35 in the tweets when compared to the blogs and +21.06 when compared to the discussion boards and *should*, which is used at a statistically less frequent rate in the tweets than the discussion boards and emails, but at a similar rate of use when compared to the blogs and SMS messages.

	Tweets						**Blogs**		
	can	***Could***	***will***	***would***	***May***	***might***	***can***	***could***	***shall***
Blogs	-11.31	-14.64	-73.16	-26.34	**-6.46**	-14.94	N/A	N/A	N/A
Discussion Boards	-93.16	-52.26	-16.31	-263.2	-19.47	-8.91	-39.66	-11.88	**1.43**
Emails	-509.32	-133.9	-689.6	-255.7	-24.43	**-6.47**	-383.9	-69.1	-20.58
SMS	-78.96	-48.33	-237	-10.71	-44.83	-31.59	-39.1	-15.75	-38.74
Tweets	N/A	N/A	N/A	N/A	N/A	N/A	11.31	14.67	-12.35

Table 5: LL comparisons of forms of modal verb use in the twitter and blog data compared to the other data-types in CANELC.

Table 5 also indicates that there is a less frequent use of the modals *can*, *could* and *shall* in the blog data than across all other text types (other patterns of use in the blogs compared to the other text-types were less prominent than this, with a high degree of fluctuation/variation from one to the next).

5.3. Summary of findings

The findings indicated that:

1. Modal verbs were used more frequently across the entire CANELC corpus than the spoken and written BNC, and more frequently in the spoken than written BNC.

2. Modals were most frequently used in the email, SMS and discussion board data respectively with, again, the rate of use in these text-types proving to be more frequent that the spoken and written BNC.
3. Modals were used at a significantly less frequent rate in tweets than the other data types (including the spoken and written BNC).
4. Modal verbs were used at a similar rate, overall, in the blogs and written BNC data, although significant differences in use of specific forms of these phenomena existed across these text-types.
5. *Can*, *will* and *would* were the most frequent core modal forms in CANELC, with *shall*, *must*, *may* and *might* proving least frequent.
6. The email and SMS data had a tendency to use most forms of modals than other data-types, with *can*, *could*, *will*, *should* and *would* proving particularly prominent in the former and *might*, *may*, *shall* and *will* in the latter. *Must* and *shall* are also used at a particularly higher rate in the written BNC than compared to the other data-types.
7. The forms *can* (which is particularly infrequent in the written BNC), *could*, *will* (which is particularly infrequent in the spoken BNC), *would*, *may* and *might* were particularly infrequent in the twitter sub-corpus when compared to the other sub-corpora. They were also less frequent, in general, in blogs, with *can*, *could*, *shall*, *would* and *should* proving particularly infrequent in this sub-corpus.

6. Discussion

The frequent use of modal forms in e-language suggests that there is a closer alignment with this 'genre' of discourse and speech, rather than communication at the written end of Crystal's continuum (although CANELC perhaps contains levels of modality, a key indicator of the 'spokenness' of discourse, that eclipse even the spoken BNC - finding 1). This is highly variable from one data-type to the next though, with the blog and twitter data perhaps aligning more closely with more formal, written discourse, and the discussion board data, emails and SMS aligning more closely to informal, spoken discourse (findings 2, 3 and 4). These result mirrors those seen in Tagg's analysis of SMS messages (2009) and Knight et al.'s analysis of specific parts-of-speech and formality in CANELC (2014).

These differences can somewhat be attributed to structural differences between blogs and tweets and the other forms in CANELC. These forms are generally outward facing so can be accessed and read by all, rather than being targeted at a more specific readership, often an individual or small group of people as with emails, SMS messages and discussion board threads. They, therefore, are not constructed with specific individuals in mind (aside from someone's twitter followers perhaps) so while it is arguably important that communication sent via these means is polite and not face-threatening, it is perhaps not *as* essential as with the other forms (this will also account for finding 7). For emails, SMS messages and discussion board threads, the frequent use of modal verbs acts as a relationship maintenance device, illustrating a 'connectedness'; a certain level of intimacy between the sender and sendee, despite the physical or temporal distance that may exist (depending on where and when it was sent and received – mirroring results seen in Knight et al., 2014; Tagg, 2009 and Herring, 2002). Also, for data-types that are likely to be communicated between known individuals (i.e. so work colleagues, partners, family members and friends), such as emails and SMS messages, relationships are particularly vital. An example of this, taken from the SMS sub-corpus, is seen below:

> ..bit early for u...Let me no what time u can be ready for x
> r only 19 quid if u book now n i think u can use clubcard vouchers 4 drayton. X
> from work and shit loads of work to do. Can u send my apologies tonight, it appears 60hr

	Hey, so u	can turn lol. Thats a good start. Its much better…
No, not yet :-s will c her in half hour... I		can really put myself into awkward pickles…
	That might b nice. Can	we speak tomorro. Give me a call in the…

Figure 4: Sample concordance output illustrating the use of *shall* in the SMS sub-corpus.

Here we see examples of where the core modal verb *can* is used as a form of a request for action or a form of negotiation between two people (lines 1, 3 and 5). It helps to soften the request put forward, again functioning as a face-saving and relationship maintenance device (making it highly interpersonal).

In terms of particular modal forms, the infrequent use of *shall*, *may* and *must* (findings 5 and 6) perhaps reconfirms that has been seen in previous research, with a general decline of the use of these forms over time (see section 2.2). Interestingly, *shall* and *must* are most frequent in perhaps the most personal form of e-language, SMS messages, despite being traditionally aligned with more formal forms of discourse, although their use is still not as prevalent as other forms are. Examples of the use of *shall* in SMS are shown in the concordance output in Figure 50 below:

...have your other christmas present :-)	Shall	i pick u up at 8?
Hehe i told her :o) when	shall	i come over then?x
...chicken. Thanks for the recipe book,	shall	hav to whip somethin up for u from one of..
Yeah,some good some less so!	Shall	return a few items and try and pick up some..
Of course.	shall	i pick u up in 45? X
We've been sent home :-) you off?	Shall	we take dog out? X

Figure 5: Sample concordance output illustrating the use of *shall* in the SMS sub-corpus.

As can be seen from these examples, the use of *shall* is highly interpersonal in the SMS data, often with an organising function, specifically when questioning and/or negotiating when the recipient, developing a sense of obligation to the request. *Shall* is often used when making plans and arranging meetings, as seen in concordance lines 1, 2, 5 and 6 here. In other instances it is used to create a sense of subjectivity about the assertions being made: s*hall hav to whip something up*; *Shall return a few items*.

A prevalence of the use of *can* and *would* on the other hand, make the e-language data generally more akin to spoken than written discourse (see Biber et al., 1999 and O'Keeffe et al., 2007), although, again, this is variable from one data-type to the next. Interestingly, the most frequent use of these forms was in the email data, which is comprised of business communication, so taken from business contexts, in CANELC rather than more informal contexts which is what might typically be expected, and was seen in previous research.

So generally, speaking why the increased use of modals when compared to spoken English (a result which conflicts the hypothesis set out at the start of this chapter)? It seems that despite being near-immediate, highly interpersonal and semi-synchronous (see section 3), forms of e-language have one key inadequacy compared to spoken discourse: the provision for effectively communicating 'beyond the word'. In face-to-face interaction we are able to access a variety of gestural, paralinguistic and extra-linguistic cues which work with spoken language to generate meaning in communication (see McNeill, 1992 and Kendon 1994). So while the words spoken are important, they are not always fully responsible, for what is being conveyed or understood (Applbaum suggests that only 35% of meaning is generated by words in communication – 1979). While features, such as contextual cues (e.g. time and location, which are often automatically recorded) and emoticons, for example, may go some way to allow for this (see Park et al., 2014 and Kalman and Gergle, 2014), we are somewhat more reliant on what is being said rather than how it is communicated when communicating via digital means.

In effect, users perhaps over-compensate for this limitation, instead relying on the language alone to build and maintain relationships; to ensure that discourse is polite and non-face-threating, making linguistic devices that function in an interpersonal way, such as modal verb forms, more frequent here than spoken and written forms.

7. Conclusion

This paper has provided an outline of the characteristics of modal verb use in various forms of e-language (across specific data-types and genres) and the relationship to their use in both spoken and written language as a 'whole'.

A key limitation of the work presented here relates to the age/currency of the data under study. The BNC, for example, is more than 20 years old so the reliability of the data in truly representing 'current day' language may be questioned. It would, therefore, be useful to replicate this study in the future once, for example, the updated spoken component of the BNC is released[6]. CANELC was also collected in 2005-2011, so this criticism may also be extended to this corpus too (and is an on-going criticism for any fixed, rather than monitor corpora). Adding to this, the representativeness and scalability of the data in CANELC can also be questioned, as it is a fairly small-sized corpus, one which includes limited samples of data across each type, genre and topic (and the prevalence of business rather than more social emails in this data type), so there is potential for extending this study in light of these shortcomings.

In addition, a more qualitative, screen by screen study of the data would also allow for the closer examination of the specific functions of the modal verb forms analysed here. A closer observation of their use between specific contributors may also help us to create a clearer profile of use across the different text-types. Finally, a focus on a wider range of forms modal verbs and a clearer distinction between the individual functions of forms, in specific contexts, would be welcomed.

However, this chapter does help to shed some light on the existence and importance of core modal verbs use in e-based communication, providing some foundational understanding of some of the characteristics and intricacies e-language as a whole. This work opens the door to a variety of interesting questions about the use of language in digital contexts, questions that can be further explored in the future.

References

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In Granger, S., Hung, J. and Petch-Tyson, S. (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: Benjamins. pp. 55-76.

Baron, N. (1998). Writing in the age of email: the impact of ideology versus technology. *Visible Language 32* (1): 35–53.

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

Benincà, P. and Poletto, C. (1997). The diachronic development of a modal verb of necessity. In van Kemenade, A-V. and Vincent, N. (Eds.), *Parameters of Morphosyntactic Change.* Cambridge: Cambridge University Press.

Berglund, Y. (1999). Utilising present-day English corpora: A case study concerning expressions of future. *ICAME Journal 24*: 25-63.

Biber, D. and Conrad, S. (1999). Lexical bundles in conversation and academic prose. In Hasselgard, H. and Okesfjell, S. (Eds.), Out of corpora: studies in honour of Stig Johansson. Amsterdam: Rodopi. pp. 181-190.

[6] A project currently underway involving researchers CASS (Centre for Corpus Approaches to Social Science) at Lancaster University (http://cass.lancs.ac.uk/) and Cambridge University Press.

Bowie, J., Wallis, S. and Aarts, B. (2013). Contemporary change in modal usage in spoken British English: mapping the impact of genre. In Arrese, J.I.M., Carretero, M., Hita, J.A. and van der Auwera, J. (Eds.), *English modality: core, periphery and evidentiality*. Berlin; Boston: De Gruyter. pp. 57-94.

Brown, P. and Levinson, S.C. (1978). Universals in language usage: politeness phenomena. In Goody, E.N. (Ed.), *Question and politeness*. Cambridge: Cambridge University Press.

Carretero, M. (1992). The role of epistemic modality in English politeness strategies. *Miscelánea 13*: 17–35.

Carter, R. and McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Coates, J., (1995). The expression of root and epistemic possibility in English. In: Aarts, B., Meyer, C. (Eds.), *The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press. pp. 145–156.

Condon, S., and Cech, C. (1996). Functional comparison of face-to-face and computer-mediated decision-making interactions. In Herring, S. (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*. Philadelphia: John Benjamins. pp. 65–80.

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

Crystal, D. (2008). *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.

Farr, F., Murphy, B. and O'Keeffe, A. (2004) The Limerick Corpus of Irish English: design, description and application. *Teanga 21*: 25 -29.

.

Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*. London: Edward Arnold.

Hansen, B. and de Haan, F. (Eds.). (2009). *Modals in the Languages of Europe. A Reference Work - Empirical Approaches to Language Typology 44*. Berlin: Mouton de Gruyter.

Herring, S. (1999). Interactional coherence in CMC. *Journal of Computer-Mediated Communication 4*(4).

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

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

Hewings, A. and Hewings, M. (2004). Impersonalising stance: a study of anticipatory 'it' in student and published academic writing. In Coffin, C. and O'Halloran, K. (Eds.), *Applying English grammar – functional and corpus approaches*. pp. 101–116.

Hinkel, E. (1995). The Use of Modal Verbs as a Reflection of Cultural Values. *TESOL Quarterly 29*: 325–343

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic essays. *Journal of Pragmatics 41*: 667-683.

Hyland, K., (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics 17*: 433–454.

Kalman, Y.M. and Gergle, D. (2014). Letter repetitions in computer-mediated communication: a unique link between spoken and online language. *Computers in human behaviour 34*: 187-193.

Kendon. A. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction 27*(3):175-200.

Kiefer, F. (1987). On defining modality. *Folia Linguistica 21*(1): 67–94.

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

Knight, D., Adolphs, S. and Carter, R. (2013). Formality in digital discourse – A study of hedging in CANELC. In Romero-Trillo, J. (Ed.) *Yearbook of Corpus Linguistics and Pragmatics.* London: Springer. pp. 131-152.

Knight, D., Adolphs, S. and Carter, R. (2014). CANELC – constructing an e-language corpus. *Corpora 9*(1): 29-56.

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

Krug, M. (2000). *Emerging English Modals: A Corpus-Based Study of Grammaticalization*. New York: Mouton de Gruyter.

Leech, G. (2011). The modals ARE declining – reply to Millar's 'Modal verbs in TIME: Frequency changes 1923-2006'. *International Journal of Corpus Linguistics 16*(4): 547-564.

Leech, G., Hundt, M., Mair, C. and Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago: The University of Chicago Press.

Millar, N. (2009). Modal verbs in TIME: Frequency changes 1923-2006. *International Journal of Corpus Linguistics 14*(2): 191-220.

Narrog, H. 2005. On defining modality again. *Language Sciences 27:* 165-192.

Nuyts, J. (2006). Modality: Overview and linguistic issues. In Frawley,W. (Ed.), *The Expression of Modality*. Berlin: Mouton de Gruyter. Berlin. pp. 1–26.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007). *From corpus to classroom: language use and language teaching.* Cambridge: Cambridge University Press.

Palmer, F. R. (1979). *Modality and the English modals*. London: Longman.

Park, J., Baek, Y.M. and Cha, M. (2014). Cross-cultural comparison of nonverbal cues in emoticons on Twitter: evidence from big data analysis. *Journal of communication 64*(2): 333-354.

Portner, P. (2009). *Modality*. Oxford: Oxford University Press

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

Schler, J., M. Koppel, S. Argamon and J. Pennebaker. (2006). Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.Shortis, T. (2007). Gr8 txtpectations: the creativity of text spelling. *English Drama Media Journal 8*: 21–6.

Sutherland, J. (2002). 'Cn u txt?'. Featured in *The Guardian*, 11 November.

Sweetser, E. E. (1982). *Root and epistemic modality: Causality in two worlds*. *Berkeley Linguistic Papers 8:* 484-507.

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

Thurlow, C., Lengel, L. and Tomic, A. (2004). *Computer mediated communication: social interaction and the internet.* London: Sage.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research 23:* 3-43.